COMPARISONS AMONG TREATMENT MEANS IN AN ANALYSIS OF VARIANCE

AGRICULTURAL RESEARCH SERVICE

OF UNITED STATES DEPARTMENT OF AGRICULTURE

HEADQUARTERS

FOREWORD

That the analysis of variance is a powerful technique for testing hypotheses has been accepted for many years. In analyzing a set of data, however, the scientist usually is interested in relationships between the means to which the analysis of variance is insensitive.

As early as 1939, statisticians used techniques independent of the analysis of variance to compare means from a given experiment. Since the middle 1950's, the interest and literature have increased almost exponentially.

In May 1957, Biometrical Services issued ARS 20-3, Mean Separation by the Functional Analysis of Variance and Multiple Comparisons. This publication has been out of print for many years. Since the publication of ARS 20-3 much work has been done on the subject, indicating the need for a major revision.

Since the job of coordinating the national aspects of statistical consulting in ARS was delegated to the Data Systems Application Division (DSAD), we asked Victor Chew, mathematical statistician, to revise ARS 20-3. We feel that he has done a very thorough job, which should put mean separation techniques in the appropriate field of reference with respect to other statistical techniques that may be used in drawing judgements from data.

Copies of this publication may be obtained from Victor Chew, University of Florida, Room 217, Rolfs Hall, Gainesville, Florida 32611.

Judson U. McGuire, Jr.
Staff Specialist, DSAD—ARS

PREFACE

The equality of the true average responses of two treatments (varieties, insecticides, concentrations, temperatures, etc.) usually is tested statistically by the Student's t-test. This is generalized for t (three or more) treatments by the F-test or the analysis of variance. If the F-test rejects the hypothesis that the t treatment means are equal, the only conclusion is that the t means are not all equal. It does not necessarily follow that these t means are all unequal although this may well be true. The next stage in the data analysis is to determine which treatment means are different. Repeated application of the Student's t-test to all possible pairs of treatment means (using pooled error either from all t samples or only from the two samples involved in the t-test) usually is discouraged since this procedure gives a large probability of getting one or more false positives (that is, of declaring two treatment means to be different, when they are, in fact, equal). Special techniques (called multiple comparison procedures) are available for this purpose.

Uses and abuses of multiple comparison procedures are discussed in this publication. One glaring abuse is its use in comparing several levels of a quantitative factor (such as concentration, temperature, and pH). Regression analysis is the appropriate technique here. Equivalently, the treatment sum of squares in the analysis of variance table should be partitioned into linear, quadratic, etc., components. In comparing the effects of, say, 10, 20, 30, and 40 p/m of a certain chemical, if the regression of the response on concentration or if any component of the sum of squares for concentrations is significant, then no multiple comparison procedure is necessary. *ALL* concentrations are significantly different in their effects. In fact, not only will 10 and 20 p/m be different, but so also will 10 and 10.1 p/m. The difference, of course, between the effects of 10 p/m and 10.1 p/m will be extremely small. However, the usual statistical test of significance is not concerned with the magnitude of the difference, but only whether a true difference exists, no matter how small.

Washington, D.C. Issued October 1977

CONTENTS

COMPARISONS AMONG TREATMENT MEANS IN AN ANALYSIS OF VARIANCE

By Victor Chew[1]

CHAPTER 1. *INTRODUCTION*

Before embarking on an experimental project, the research scientist should carefully consider various issues. These issues include questions that the experiment hopefully will answer, the factors or variables to be controlled or kept constant during the experiment, the levels of the factors to be varied in the study, the number of observations to be taken, and the manner in which these observations will be grouped into blocks. We shall need fewer observations or have wider applicability of the results, or both, if the experiment is designed efficiently.

This publication is concerned with a particular facet of the analysis of the experimental data, assuming that the experiment has been designed properly. It is applicable irrespective of the experimental design (completely randomized, randomized blocks, Latin square, split plot, etc.). We also shall assume that the reader is familiar with the computational aspects of the analysis of variance for these designs.

The basic terms and notions in statistical inference will be reviewed in this chapter. This is necessary to understand the relative merits of multiple comparison procedures that are currently available.

In the simplest *hypothesis testing* situation, we compare two *treatments* (varieties of peanuts, fertilizers, temperatures, pH, machine settings, etc.). If we denote the true means of the two treatments by μ_1 and μ_2, the statistical hypothesis to be tested is usually that these two means are equal ($\mu_1 = \mu_2$). This hypothesis, called the *null hypothesis*, often is denoted by H_0. We write it as H_0: $(\mu_1 - \mu_2) = 0$. (We can test a more general hypothesis, viz., $(\mu_1 - \mu_2) = d$, where d is specified numerically.)

In classical hypothesis testing, we must decide whether to accept or to reject H_0. (In sequential testing, we allow a third alternative of requiring more observations to be taken.) Because the true or *population means* μ_1 and μ_2 are unknown and unknowable, our decision from the statistical test (whether to accept or reject H_0) is subject to error. If $\bar{y}_1$ and $\bar{y}_2$ are the observed or *sample* means, estimating μ_1 and μ_2 respectively, then because of nonhomogeneity of the experimental material (such as plants, animals, plots of land, batches of peanuts), failure to reproduce identical experimental conditions, errors of measurements, etc., $\bar{y}_1$ and $\bar{y}_2$ will be unequal, even if μ_1 and μ_2 are equal. In fact, we may even have $\bar{y}_1$ larger than $\bar{y}_2$ when actually μ_1 is smaller than μ_2, especially from a small experiment.

There are two kinds of error in hypothesis testing:

Type I—Reject H_0 when H_0 is, in fact, true (i.e., erroneously deciding that μ_1 and μ_2 are unequal).
Type II—Accept H_0 when H_0 is, in fact, false (i.e., incorrectly deciding that μ_1 and μ_2 are equal).

The probabilities of a test making these errors usually are denoted by α and β, respectively. The perfect test is, of course, infallible (where $\alpha = \beta = 0$), but this is impossible with a finite sample. A good experiment is one in which both α and β are small. The value of α is called the *significance level* of the test, sometimes expressed as a percentage. By suitably choosing the *rejection region* or *critical values* for the test statistic, we can make α as small as we like, but only at the expense of increasing β. For example, we can make $\alpha = 0$ by *always* accepting H_0, regardless of the experimental data, but in this case $\beta = 1$. The only way to decrease both α and β simultaneously is to increase the sample size (number of observations). Conventionally, α is taken to be equal to .05 or .01. With β defined as the probability of accepting H_0 when H_0 is false, $(1 - \beta)$ is the probability

[1] Mathematical statistician, Biometrical and Statistical Services, Agricultural Research Service, U.S. Department of Agriculture, 217 Rolfs Hall, University of Florida, Gainesville, Fla. 32611

of rejecting H_0 when H_0 is false. This quantity is called the *power* of the test—the probability of the test to detect a difference when one exists. There are infinitely many tests with the same value of α; among these, we choose the most powerful one (for which β is least) if one exists.

If H_0 is false, another *alternative hypothesis* (denoted by H_a) is true. Corresponding to H_0: $(\mu_1 - \mu_2) = 0$, three possible alternative hypotheses are $(\mu_1 - \mu_2) > 0$, $(\mu_1 - \mu_2) < 0$, and $(\mu_1 - \mu_2) \neq 0$, called the right-tail, left-tail, and two-tail alternative hypotheses, respectively. If the first treatment is "control" (i.e., no treatment at all), the second treatment is the application of some insecticide, and the response being measured is the number of a particular insect per plant, we know *a priori* that the alternative to H_0:$\mu_1 = \mu_2$ is H_a: $\mu_1 > \mu_2$ because the application of the insecticide cannot possibly *increase* the average count. By capitalizing on the one-sidedness of H_a, we can construct a more powerful test of H_0, with the same α. If we are comparing two new insecticides, the alternative hypothesis is two-sided.

It will be seen that α is associated with H_0 and β with H_a. This explains why we can control α but not β. We need the actual difference between the two means to control β. For this reason, experimenters too often ignore Type II errors. If they are only concerned with holding Type I errors down to 5%, they need not conduct the experiment at all. They merely need to take 20 index cards, mark one with an X, shuffle them thoroughly, and draw one card at random. Reject H_0 if the marked card is drawn. At a saving of hundreds if not thousands of dollars, this experimenter has only a 5% chance of making a Type I error. The reader should think about the value of β in this case.

We cannot emphasize strongly enough the distinction between *statistical* and *practical* significance. Any difference between the sample means $\bar{y}_1$ and $\bar{y}_2$, *no matter how small, must* be declared statistically significant if the population or true means μ_1 and μ_2 are unequal, unless the test has committed a Type II error (incorrectly declaring two means equal). The test *will* declare the difference significant if we have enough replications. In calculating the number "n" of observations to be taken, we only should require n to be large enough so that the test will detect a difference of at least d (of practical significance) between μ_1 and μ_2. It is no big loss to declare incorrectly that μ_1 and μ_2 are equal if they differ by an insignificant amount.

The author thinks that the research worker has been oversold on hypothesis testing. Just as no two peas in a pod are identical, no two treatment means will be exactly equal. They always will be different, even if only in the thousandth decimal place. It seems ridiculous, therefore, to test a hypothesis that we *a priori* know is almost certain to be false. If the test accepts the hypothesis of equal treatments, a Type II error probably has occurred. A related but much more informative alternative approach is *interval estimation* of $(\mu_1 - \mu_2)$. The *confidence limits*, of the form $(\bar{y}_1 - \bar{y}_2) \pm c$, will tell us whether the null hypothesis will be accepted (if the limits have different signs) or rejected (if they have the same signs). They also will give the estimated magnitude of the actual difference. The value of c depends, among other things, on the *confidence level* γ. If $\gamma = 0.95$, we have 95% confidence that $(\mu_1 - \mu_2)$ is between $(\bar{y}_1 - \bar{y}_2 - c)$ and $(\bar{y}_1 - \bar{y}_2 + c)$. The closer γ is to unity, the wider the confidence interval. For a given γ, we can shorten the interval by increasing the sample size.

The practice of hypothesis testing when comparing several treatments is even more difficult to justify. When comparing 10 new varieties of corn, for example, it is inconceivable that all the true average yields will be exactly equal. Besides a simultaneous confidence interval approach for all pairs of varieties, a better objective may be to select the smallest subgroup that has a preassigned probability (95%, say) of including the highest yielding variety. This subgroup of varieties may be tested more intensively and compared in a later experiment, as in the screening of new drugs.

CHAPTER 2. *PARTITIONING OF DEGREES OF FREEDOM FOR TREATMENTS*

This chapter deals with situations in which it is possible, *before* performing the experiment, to partition the degrees of freedom (d.f.) for treatments, either completely into single d.f. or partially into groups of d.f. Partitioning must not be suggested *after* examination of the experimental data. LeClerg *(1957)* [2] referred to this partitioning as "functional analysis of variance." *Use of a multiple comparison procedure in this chapter (with a couple of exceptions, explicitly stated) constitutes an abuse of the technique.* If the difference between

[2] The year in parentheses following the author's name refers to List of References, p. 59.

the observed average responses of two treatments is statistically significant, we shall simply say that the two treatments are different.

In this chapter, a significant F-test for treatments is *not* a prerequisite for the partitioning of the treatments d.f. or s.s. (sum of squares). In fact, the F-test need not and should not be carried out at all. In comparing t treatments, with $(t-1)$ d.f., the blanket or overall F-test for treatments is averaged over $(t-1)$ orthogonal comparisons (defined later). If only one or two of these comparisons (or contrasts) are significant, the overall F-test is diluted or weakened by the $(t-2)$ or $(t-3)$ nonsignificant contrasts and erroneously may give a nonsignificant F value.

2.1 *Orthogonal Contrasts*

Let $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_t$ and $T_1, T_2, \ldots, T_t$ be the sample means and totals from Treatments 1, 2, . . . , t, respectively. (Unless otherwise stated, we shall assume that the treatments are equally replicated. If n is the common number of replicates per treatment, we have $\bar{y}_i = T_i/n$.) The expression $\Sigma a\bar{y} = (a_1\bar{y}_1 + \ldots + a_t\bar{y}_t)$ is called a linear combination of the treatment means. A linear combination is called a *comparison* or a *contrast* if the coefficients (the a's) add up to zero. For example, if we have $t = 4$ treatments, $\bar{y}_1 - (\bar{y}_2 + \bar{y}_3 + \bar{y}_4)$ is a linear combination of the treatment means. It is not a contrast, however, since the sum of coefficients is nonzero. (It is equal to -2.) This linear combination compares the mean of the first treatment with the *sum* of the means of the remaining three treatments, which is not a fair comparison according to the ordinary meaning of "fair." A fair comparison is to compare $\bar{y}_1$ with the *average* of the means of the remaining three treatments, given by $y_1 - (\bar{y}_2 + \bar{y}_3 + \bar{y}_4)/3$, which is now also a contrast since the coefficients add up to zero. To avoid fractional coefficients, the preceding contrast usually is written $3\bar{y}_1 - (\bar{y}_2 + \bar{y}_3 + \bar{y}_4)$.

The sum of squares corresponding to a contrast $C = \Sigma a\bar{y}$ is

$$\text{s.s. } (C) = n(\Sigma a\bar{y})^2/(\Sigma a^2) = (\Sigma aT)^2/[n(\Sigma a^2)], \tag{2.1}$$

where Σa^2 is the sum of the squares of the coefficients in the contrast. (Notice that the s.s. is unchanged if we multiply the coefficients by a constant.) Since a contrast has one d.f., the s.s. is also a mean square (m.s.) because (m.s.) = (s.s.)/(d.f.). It may be tested for significance by dividing it by the error m.s. (with m d.f., say) that normally would be used to make the overall test for treatments in the analysis of variance. The calculated ratio is compared with the critical value of the F-distribution with 1 and m d.f.

If we are comparing $t = 4$ treatments in a completely randomized experiment with $n = 3$ replicates per treatment, the d.f. for the error m.s. is $m = t(n-1) = 8$. In a 5% two-tail test, the critical value of the F-distribution with 1 and 8 d.f. is 5.32. If a one-tail test is justifiable (as, for example, if in the contrast $3\bar{y}_1 - (\bar{y}_2 + \bar{y}_3 + \bar{y}_4)$, the first treatment is control and the other treatments are three types of insecticides), the 5% critical value is only 3.46. Since a smaller critical value is easier to exceed, a significant difference is easier to declare in a one-tail test. Consequently, the test is less likely to commit a Type II error (failure to declare a difference when one exists).

Two contrasts, $C_1 = \Sigma a\bar{y}$ and $C_2 = \Sigma b\bar{y}$, are said to be *orthogonal* if $\Sigma ab = 0$ (i.e., if the sum of the products of the corresponding coefficients in the two contrasts is zero). A set of contrasts is said to be mutually orthogonal if all pairs of contrasts in the set are orthogonal. If, for brevity, we write $(a_1\bar{y}_1 + a_2\bar{y}_2 + \ldots + a_t\bar{y}_t)$ as $(a_1, a_2, \ldots, a_t)$, the three contrasts $(1, 1, -1, -1)$, $(1, -1, -1, 1)$, and $(1, -1, 1, -1)$ are mutually orthogonal. It can be proved that there are only $(t-1)$ mutually orthogonal contrasts among t means; however, there are infinitely many such sets of mutually orthogonal contrasts.

The following are another two sets of mutually orthogonal contrasts: $(1, 1, -1, -1)$, $(1, -1, 0, 0)$, $(0, 0, 1, -1)$, and $(3, -1, -1, -1)$, $(0, 2, -1, -1)$, $(0, 0, 1, -1)$. It also can be proved that if $C_1, C_2, \ldots, C_{t-1}$ are $(t-1)$ mutually orthogonal contrasts, their individual sums of squares add up exactly to the treatments s.s. The statistical distributions of these contrasts are independent. This is one reason why, whenever possible, we should aim for an orthogonal decomposition of the treatments d.f. Of the possible sets of mutually orthogonal contrasts, the experimenter should choose the set that is most interesting or most relevant to his study. Mutual orthogonality is desirable but not absolutely essential. If several contrasts interest the scientist, he should not let the lack of mutual orthogonality prevent him from performing the statistical tests, as long as these contrasts have not been suggested by the data. Contrasts suggested after data snooping should be tested by a multiple comparison procedure.

2.2. *Qualitative Factors*

Experimental variables or factors may be divided into qualitative and quantitative factors. Examples of qualitative factors are varieties (peanuts, corn, etc.), types (soils, fungicides, etc.), locations, and methods of chemical analyses or of counting bacteria. Examples of quantitative factors are temperature, pressure, humidity, pH, concentration, and several levels of a fertilizer. Although the various varieties or soil types in an experiment also are referred to as the levels of the factors "varieties" and "soil types," no meaningful numerical values can be assigned to the levels of a qualitative factor. Levels of a quantitative variable are, of course, naturally numerical.

Factorial experiments are those in which the treatments are made up of all possible combinations of the levels of two or more factors (qualitative or quantitative). (The term "factorial" thus merely describes the nature of the treatments and not the design of the experiment, which may be completely randomized, randomized block, Latin square, split-plot, etc.) The simplest factorial is the 2^2 or 2 x 2 experiment, with two factors A and B, each at two levels. For the 2 x 2 factorial, the partitioning of the d.f. for treatments is the same whether the two factors are both qualitative or quantitative, or one of each kind. The two levels may be designated generally as H (high) or L (low). The low level, in particular, may be zero. For a qualitative factor, we may arbitrarily label one level H and the other L. The four treatments are denoted by (1), a, b, and ab, where absence of a letter implies that the corresponding factor is at the low level; and (1) is a special symbol for the treatment where both factors are at the low level. These four treatments could have been more explicitly but awkardly denoted by A_LB_L, A_HB_L, A_LB_H, and A_HB_H, respectively.

The three d.f. for treatments are partitioned into the main effect of A, main effect of B, and their interaction. The coefficients for these contrasts are as follows:

	Treatments				
Contrasts	(1)	a	b	ab	
C_1	−1	1	−1	1	Main effect of A
C_2	−1	−1	1	1	Main effect of B
C_3	1	−1	−1	1	Interaction of A and B

The coefficients for the main effect of A are +1 for treatments where A is at the high level and −1 if A is at the low level; and similarly for B. The coefficients for interaction are obtained by multiplying corresponding coefficients for main effects. To get the sums of squares for the preceding contrasts, we apply Equation (2.1) to the four treatment means or totals, using the coefficients for each contrast in turn.

The difference [a − (1)] is called the *simple* effect of A at the low level of B; similarly, (ab − b) is the simple effect of A at the high level of B. The main effect of A is the average of the simple effects of A. (To avoid fractions, the coefficients for this average have been multiplied by two. The reader will recall that s.s. for a contrast is unchanged if the coefficients are multiplied by a common number.)

If the factors A and B act independently, the two simple effects of A should be about the same. (Experimental or random errors will prevent them from being exactly equal.) Therefore, their difference

$$(ab - b) - [a - (1)] = ab + (1) - a - b = C_3$$

should be approximately zero if A and B are independent. If this quantity is large (significantly different from zero), we say that there is interaction between A and B (i.e., effect of A at low level of B is different from effect of A at high level of B). We also can write C_3 as (ab − a) − [b − (1)] = (effect of B at high level of A) − (effect of B at low level of A) so that if effect of A depends on the level of B, we know that the effect of B depends on the level of A.

The following artificial two-way tables of means show some possible results of the tests for main effects and interaction. In (d), for example, the simple effect of A is 10 units at low B and 20 units at high B, showing dependence of the effect of A on the level of B or interaction between A and B.

		A		
		Low	High	Average
B	Low	10	20	15
	High	12	24	18
	Average	11	22	

(a)

A sig.
B not sig.
A x B not sig.

		A		
		Low	High	Average
B	Low	10	20	15
	High	22	34	28
	Average	16	27	

(b)

A sig.
B sig.
A x B not sig.

		A		
		Low	High	Average
B	Low	10	20	15
	High	6	26	16
	Average	8	23	

(c)

A sig.
B not sig.
A x B sig.

		A		
		Low	High	Average
B	Low	10	20	15
	High	18	38	28
	Average	14	29	

(d)

A sig.
B sig.
A x B sig.

In general, a two-factor experiment is a p $\times$ q factorial. The (pq $-$ 1) d.f. for treatments will be partitioned into main effects of A with (p $-$ 1)d.f., main effects of B with (q $-$ 1) d.f., and interaction with (p $-$ 1) (q $-$ 1) d.f. The A $\times$ B interaction is more difficult to illustrate if p and q are greater than two, but the interpretation is similar to that in the 2 $\times$ 2 factorial; viz., differences among levels of A depend on the levels of B, and vice versa. If the p levels of A are such that orthogonal contrasts are possible, the (p $-$ 1) d.f. for the main effects of A should be partitioned further into single d.f. If it is impossible to partition the (p $-$ 1) d.f. for A, then it is legitimate to use a multiple comparison procedure to compare the p levels of A.

Testing the main effects of A presupposes that there is no A $\times$ B interaction. If interaction exists, the differences among the levels of A depend on the level of B. It does not make much sense to compare the levels of A *averaged* over all levels of B, which is what main effect is. It is more instructive to compare the levels of A

for *each* level of B separately, and vice versa, using the pooled error mean square from the complete experiment, if the assumption of homogeneous variances is valid.

With three factors, the simplest is a 2^3 or $2 \times 2 \times 2$ factorial. The eight treatments may be denoted by (1), a, b, ab, c, ac, bc, abc, in an obvious extension of the previous notation, where, for example, ac stands for the treatment with factors A and C at their high level and B at the low level. The seven d.f. for treatments will be partitioned into main effects (A, B, C), two-factor (or first order) interactions (A × B, A × C, B × C), and three-factor (or second order) interaction (A × B × C), each with a single d.f. Second and higher order interactions are difficult to interpret. The A × B × C interaction is the interaction of (A × B) and C. If A × B × C interaction is significant, the A × B interaction at the high level of C is different from that at the low level of C. The coefficients for the following contrasts are obtained as in the 2 × 2 factorial experiment.

	Treatments							
	(1)	a	b	ab	c	ac	bc	abc
A	−1	1	−1	1	−1	1	−1	1
B	−1	−1	1	1	−1	−1	1	1
A × B	1	−1	−1	1	1	−1	−1	1
C	−1	−1	−1	−1	1	1	1	1
A × C	1	−1	1	−1	−1	1	−1	1
B × C	1	1	−1	−1	−1	−1	1	1
A × B × C	−1	1	1	−1	1	−1	−1	1

The $2 \times 2 \times 2$ factorial can be generalized to the $p \times q \times r$ factorial (three factors A, B, and C, with p, q, and r levels, respectively), to the 2^p factorial (p factors, each at two levels), and to the $p_1 \times p_2 \times \ldots \times p_r$ (r factors with $p_1, p_2, \ldots, p_r$ levels). The total number of treatment combinations increases rapidly with increasing number of factors. With six factors, even if each is at two levels, we require $2^6 = 64$ experimental units per replicate. Besides the 6 main effects, there will be 15 two-factor, 20 three-factor, 15 four-factor, 6 five-factor, and 1 six-factor interactions. If we can assume that high order interactions (four-factor or higher, say) do not exist, as is usually true, we may pool these interactions for use as error mean square so that we do not need to replicate. In fact, a single replicate already may be too large an experiment, and our resources may allow us to carry out only a portion of the full factorial experiment. So-called fractional factorial experiments are available for this purpose. They are discussed in Davies (*1956*), Cochran and Cox (*1957*), Peng (*1967*), John (*1971*), and Anderson and McLean (*1974*).

The following example, taken from Little and Hills (*1972*), shows the partitioning of treatments d.f. to give meaningful single d.f. contrasts. Six sources of nitrogen on yield of sugar beet were compared: Control (1), urea (2), ammonium sulfate (3), ammonium nitrate (4), calcium nitrate (5), and sodium nitrate (6).

	Treatments						
Contrasts	1	2	3	4	5	6	
C_1	−5	1	1	1	1	1	Nitrogen vs. no nitrogen
C_2	0	−4	1	1	1	1	Organic vs. inorganic nitrogen
C_3	0	0	−1	−1	1	1	Ammonium vs. nitrate nitrogen
C_4	0	0	−1	1	0	0	Ammonium nitrate vs. sulfate
C_5	0	0	0	0	−1	1	Calcium vs. sodium nitrate

The reader should check the mutual orthogonality of the contrasts. Note that the interpretation of Contrast C_3 is not quite right since Treatment 4 contains both ammonium and nitrate nitrogen.

An interesting factorial experiment was conducted by Dr. Ralph Segall at the U.S. Horticultural Research Laboratory in Orlando, Fla. He studied the effects of 10 fertilizer treatments on the incidence of postharvest bacterial soft-rot of tomato fruits. The 10 treatments (all of which had 18–0–25) initially may be regarded as a 2×5 factorial (mulching at two levels and "additives" at five levels). The five additives are made up of control and four chemicals. The four chemicals are in the form of a 2×2 factorial (2 anions and 2 cations). We have shown the coefficients for only five mutually orthogonal contrasts. The remaining four contrasts are the interactions between C_1 and each of C_2, C_3, C_4, and C_5. The reader may interpret the contrasts $C_1, \ldots, C_5$ and the interactions between C_1 and each of $C_2, \ldots, C_5$.

	Treatments		Contrasts				
			C_1	C_2	C_3	C_4	C_5
Mulched beds	Control	(1)	1	−4	0	0	0
	Calcium nitrate	(2)	1	1	1	1	1
	Calcium chloride	(3)	1	1	1	−1	−1
	Potassium nitrate	(4)	1	1	−1	1	−1
	Potassium chloride	(5)	1	1	−1	−1	1
Nonmulched beds	Control	(6)	−1	−4	0	0	0
	Calcium nitrate	(7)	−1	1	1	1	1
	Calcium chloride	(8)	−1	1	1	−1	−1
	Potassium nitrate	(9)	−1	1	−1	1	−1
	Potassium chloride	(10)	−1	1	−1	−1	1

There may be situations in which it is justifiable to apply a multiple comparison procedure to compare factorial treatments. For example, suppose a farmer is interested in growing one of three types of grasses and using one of four types of fertilizers. The farmer is not interested in the scientific comparison of yields from the three varieties of grasses or types of fertilizers. He is only interested in maximizing his profit. If the commercial values of the three grasses and the costs of the four fertilizers are different, analyzing the profit (in dollars and cents) per plot is more relevant than analyzing yields per plot. The 12 treatments (combinations of grasses and fertilizers) may be compared for profitability, using a multiple comparison procedure and ignoring their factorial nature.

At a panel discussion sponsored by the Data Systems Application Division, Agricultural Research Service, during the joint meeting of the statistical societies in Atlanta in August 1975, two panel members (Dr. David B. Duncan and Dr. John W. Tukey) said they might condone multiple comparisons of individual factorial treatments (from qualitative factors) if the main effects were not significant (Duncan) or if their F ratios were less than two (Tukey).

2.3. *Quantitative Factors*

2.3.1. *One Factor*

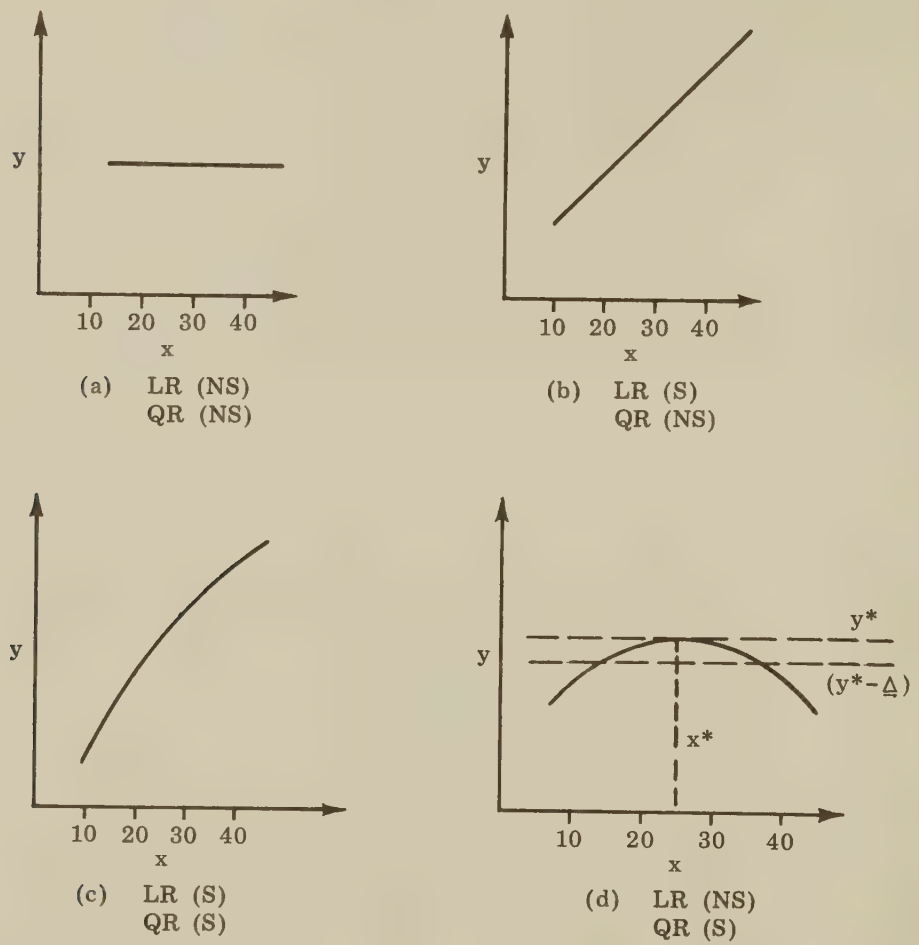
With a quantitative factor (e.g., temperature, pressure, humidity, pH, and concentration or levels of a fertilizer), regression analysis or curve fitting is the most appropriate technique. The treatments d.f. and s.s. should be partitioned into components due to linear (first degree) regression, quadratic (second degree) regression, cubic (third degree) regression, and so forth. If enough theoretical knowledge exists to specify the mathematical form of the relationship between the response y and the experimental variable x (e.g., logistic, Mitscherlich's law, Gompertz's law, von Bertalanffy's curve, etc.), this equation should be fitted to the data. In most (if not all) agricultural experimentations, however, the mathematical relationship between the response and the so-called independent variable is so complex that it defies specification. Therefore, we must approximate the unknown mathematical relationship by means of a polynomial of the form $y = b_0 + b_1x + b_2x^2 + \ldots + b_dx^d$. Within a limited range of the independent variable, a polynomial approximation is usually satisfactory if the response does not level off in the experimental range of x, in which case an asymptotic curve should be fitted.

Table 1 shows the analysis of variance of a randomized block experiment with b replicates or blocks, t treatments (levels of a quantitative factor), and m measurements per plot (experimental unit), with partitioning of the treatments d.f. and s.s. into linear and quadratic components. With the general availability of computer programs, it is not difficult to fit a polynomial of a higher degree than quadratic. The ratio ms(dr)/ms(e) provides a test for the statistical significance of the combined contributions from the higher order polynomials, sometimes called a test of the lack of fit of the fitted model (in this case quadratic). If quadratic is sufficient, this ratio has the F-distribution with $(t - 3)$ and $(b - 1)(t - 1)$ d.f. (For testing, the author generally recommends the use of ms(e) rather than ms(s) as the error term since the latter does not represent true replications. If $b = 1$, we are forced to use ms(s) as the error term, but this is dangerous since ms(s) may seriously underestimate ms(e) and it will then be easy to get a spuriously significant result.)

If the quadratic term is statistically significant but its s.s. is only a small part of the treatments s.s., we may prefer to fit a linear trend only since the curvature of the response curve is only slight. We may be able to predict the response y better (i.e., with a smaller mean squared error of prediction) by using a straight line rather than a quadratic, even if the true response curve is a quadratic function. The curvature, however, must be slight. This comes about through having to estimate fewer parameters (constants of the response function) in linear regression. A straight line is also easier to use than a parabolic curve.

In comparing the effects of, say, 10, 20, 30, and 40 p/m of a certain chemical, if the linear or quadratic regression of response on concentration is significant, or both are significant, no multiple comparison procedure is necessary. *All* concentrations are significantly different in their effects. In fact, even 10 and 10.1 p/m also will be different. Of course, the difference between the effects of 10 and 10.1 p/m will be extremely small. The usual significance test is not concerned with the magnitude of the difference, however. It is only concerned about whether a true difference exists, no matter how small.

We have the following possible results with one factor:



LR = linear regression; QR = quadratic regression
S = significant; NS = not significant

In (a), *all* treatments (infinitely many between 10 and 40 p/m) are the same, while in (b) and (c) all treatments are different. In (d), all treatments less than x* (the value of x that will maximize y) are different. We may want to estimate x* and construct confidence limits for it. If y* is the maximum response, we may be interested in finding the range of x that will give a response higher than (y* − Δ), where (y* − Δ) is an acceptably high yield. If it costs more to apply the factor x the higher its level is, we should take z as the

Table 1. *Analysis of variance of a randomized block experiment to compare effects of several levels of a quantitative factor*

Sources of variation	d.f.	s.s.	m.s.	F
Blocks (B)	b–1	ss(b)	ms(b)	ms(b)/ms(e)
Treatments (T)	t–1	ss(t)	ms(t)	ms(t)/ms(e)
Linear regression	1	ss(ℓ r)	ms(ℓ r)	ms(ℓ r)/ms(e)
Quadratic reg. (additional)	1	ss(qr)	ms(qr)	ms(qr)/ms(e)
Deviations from reg.	t–3	ss(dr)	ms(dr)	ms(dr)/ms(e)
Error (B x T)	(b–1) (t–1)	ss(e)	ms(e)	------
Subsampling error	bt(m–1)	ss(s)	ms(s)	------
Total	btm–1	ss(T)	----	------

response variable, where z is the yield per unit cost of application of x. These considerations are more meaningful than the question often asked by the naive experimenter: Among 10, 20, 30, and 40 p/m, which are different in their effects?

There are two options if the lowest level of x in the experiment is zero (control). We may fit a regression curve to all levels (including zero), or we may isolate a single d.f. for the contrast between zero and nonzero levels and fit a regression curve to the nonzero levels only. Quite often the regression is curvilinear in the first option and linear in the second option. If this is so, the second method of analysis is preferable, especially if in actual usage the factor x will not be applied at a level below the first nonzero level of the experiment.

For the linear regression model $y = b_0 + b_1x$, the estimated responses at $x = x^*$ and at $x = x^{**}$ are $y^* = b_0 + b_1x^*$ and $y^{**} = b_0 + b_1x^{**}$, respectively. Therefore, the estimated difference in response at any two values x^* and x^{**} is equal to $b_1(x^{**} - x^*)$, and the variance of this estimated or predicted difference is $(x^{**} - x^*)^2$ (variance of b_1). The formula for the variance of b_1 is given in Equation (2.4). The $100\,(1-\alpha)\%$ confidence interval for the true difference is $b_1(x^{**} - x^*) \pm t(\alpha;\nu)\sqrt{(x^{**} - x^*)^2\,(\text{estimated variance of } b_1)}$, where $t(\alpha;\nu)$ is the two-sided $(100\,\alpha)\%$ point of Student's t-distribution with ν d.f.

For the quadratic regression model $y = b_0 + b_1x + b_2x^2$, the estimated difference is $b_1(x^{**} - x^*) + b_2(x^{**2} - x^{*2})$, with variance equal to $[(x^{**} - x^*)^2$ (variance of b_1) $+ (x^{**2} - x^{*2})^2$ (variance of b_2) $+ 2(x^{**} - x^*)(x^{**2} - x^{*2})$ (covariance of b_1 and b_2)]. In a good regression computer program, the printout will include the estimated variances and covariances of the estimated regression coefficients.

Because linear relationships occur frequently, we will give the computational results for linear regression analysis. In general, let $\bar{y}_i$ be the mean of the n_i observations taken at x_i, the i^{th} level of the factor ($i = 1, 2, \ldots, t$). (We are allowing unequal replications here. In Table 1, $n_i = bm$, a constant.) The equation of the fitted line is $\bar{y} = b_0 + b_1 x$, where

$$b_1 = \frac{\Sigma n_i x_i \bar{y}_i - (\Sigma n_i x_i)(\Sigma n_i \bar{y}_i)/N}{\Sigma n_i x_i^2 - (\Sigma n_i x_i)^2/N}, \tag{2.2}$$

$$b_0 = [(\Sigma n_i \bar{y}_i) - b_1 (\Sigma n_i x_i)]/N, \tag{2.3}$$

and $N = (n_1 + n_2 + \ldots + n_t)$, the total number of observations. (In the simplest linear regression problem, $n_1 = n_2 = \ldots = n_t = 1$, and the above formulas for the slope and intercept of the line will reduce to more familiar ones.) The s.s. for linear regression is (Num.)²/Den., where "Num." and "Den." are the numerator and denominator, respectively, of the expression for b_1 above. The s.s. for deviations from regression, now with $(t - 2)$ d.f. if we are only fitting a straight line, is most conveniently obtained by subtracting ss(ℓ r) from ss(t), the treatments s.s. Finally, the variance of b_1 is

$$\text{var. } (b_1) = \sigma^2/[\Sigma n_i x_i^2 - (\Sigma n_i x_i)^2/N], \tag{2.4}$$

and σ^2 may be estimated by ms(e) in Table 1, or by ms(dr) if $b = 1$.

If the levels are replicated equally and spaced equally, the computations for obtaining the various s.s. for regression will be simplified considerably by the use of orthogonal polynomials, shown in Table 2 for 3, 4, and 5 levels only. For more extensive tables and discussion of the method for getting the actual regression equation, see Fisher and Yates (*1963*). If we look at $t = 4$ levels, say, in Table 2, we see that the three sets of

coefficients form a set of mutually orthogonal contrasts. (A polynomial curve of degree (t − 1) will pass through the t means exactly.) With these coefficients, we can obtain the s.s. for linear or quadratic regression, using Equation (2.1) in the previous section on orthogonal contrasts. An example follows.

Table 2. *Orthogonal polynomials*
(t = number of levels; d = degree of polynomial)

t=3		t=4			t=5			
d=1	d=2	d=1	d=2	d=3	d=1	d=2	d=3	d=4
−1	+1	−3	+1	−1	−2	+2	−1	+1
0	−2	−1	−1	+3	−1	−1	+2	−4
+1	+1	+1	−1	−3	0	−2	0	+6
		+3	+1	+1	+1	−1	−2	−4
					+2	+2	+1	+1

Chew (*1962*) discussed published results of an experiment wherein the research worker erroneously concluded that there were no treatment differences, through failure to partition the treatments d.f. and s.s. Table 3 shows the analysis of variance and treatment means with b = 5 blocks, t = 4 treatments (0, 2, 4, and 6 degrees of angle), and m = 5 repeated measurements on each experimental unit. (The response was the force in pounds required to separate a set of electrical connectors at various angles of pull.) The treatment means show increasing response with increasing angles. Each treatment mean was an average of n_i = bm = 25 observations. From the coefficients in Table 2, the means in Table 3 and Equation (2.1), we have the following sums of squares for regression:

$$\text{linear regression} = \frac{25[(-3)(41.94) + (-1)(42.36) + (1)(43.82) + (3)(46.30)]^2}{(-3)^2 + (-1)^2 + (1)^2 + (3)^2} = 264.26$$

$$\text{quadratic regression} = \frac{25[(1)(41.94) + (-1)(42.36) + (-1)(43.82) + (1)(46.30)]^2}{(1)^2 + (-1)^2 + (-1)^2 + (1)^2} = 26.52$$

$$\text{cubic regression} = \frac{25[(-1)(41.94) + (3)(42.36) + (-3)(43.82) + (1)(46.30)]^2}{(-1)^2 + (3)^2 + (-3)^2 + (1)^2} = 0.01$$

In a two-tail test, the F-ratio for linear regression is significant at between the 2½% and the 1% level. In a one-tail test it will be significant at between the 1¼% and the ½% level. (A one-tail test could be justified here.)

Table 3. *Analysis of variance and means*

Source of variation	d.f.	s.s.	m.s.	
Blocks	4	1234.83	308.71	
Treatments:	3	290.79	96.93	2.56 (not sig.)
Linear regression	1	264.26	264.26	6.97*
Quadratic regression	1	26.52	26.52	<1
Cubic regression	1	.01	.01	<1
Error	12	455.03	37.92	
Subsampling error	80	316.50	3.96	
Total	99	2297.15		
x: 0	2	4	6	
$\bar{y}$: 41.94	42.36	43.82	46.30	
Difference:	0.42	1.46	2.48	

With $n_i = 25$ and $N = 100$, the formulas for the slope and intercept give:

$$b_1 = \frac{(25)[0(41.94) + 2(42.36) + 4(43.82) + 6(46.30)] - (25)(12)(25)(174.42)/100}{(25)(0 + 4 + 16 + 36) - [25(12)]^2/100}$$

$$= 0.727;$$

$b_0 = [25(174.42) - 0.727(25)(12)]/100 = 41.424,$

so that the equation is $y = 41.424 + 0.727x$.

Since regression is significant, no multiple comparisons are necessary. The treatments are *ALL* different (in their effects). For example, 0 and 2 degrees are different (without testing), as well as 0 and 1 degree or even 0 and 0.1 degree. This equation gives an estimate of y for any given x; and, clearly, for two different values of x, the equation gives different values of y. The difference in response at $x = x^*$ from that at $x = x^{**}$ is

$$y(\text{at } x^{**}) - y(\text{at } x^*) = 0.727\,(x^{**} - x^*),$$

and its estimated variance is $(x^{**} - x^*)^2\,(37.92)/\{(25)(56) - [25(12)]^2/100\} = 0.0758\,(x^{**} - x^*)^2$, using Equation (2.4) for the variance of b_1. The 95% confidence interval for the difference in the two responses corresponding to a unit difference in the x values is $0.727 \pm 2.179\sqrt{.0758} = 0.727 \pm .600 = (.127, 1.327)$.

If the observed means of the t levels are in increasing (or decreasing) order and t is at least four, no further statistical test is necessary to establish significance of treatment effects, if it is known *a priori* that the effect of treatment, if any, is to increase (or decrease) the response, for the probability of the t means falling in that order under the null hypothesis is $1/(t!) \leq 1/24$, if $t \geq 4$, which is significant at the conventional 5% level. If there is no prior knowledge of the direction of the treatment effect, a two-sided test is necessary and t has to be at least five for the ordering of the t means to be significant at the 5% level.

For a criticism of the widespread misuse of Duncan's multiple range test in agricultural research to compare levels of a quantitative factor, see Mead and Pike (*1975*), particularly Section 2.2.

2.3.2. *Two or More Factors*

For one quantitative factor, we partition the treatments d.f. into linear, quadratic, cubic, etc., regression, which is equivalent to fitting a polynomial of the form $y = b_0 + b_1x + b_2x^2 + \ldots + b_dx^d$, where y is the measured response and x is the level of the experimental factor. We similarly analyze two quantitative factors A and B. Denote the levels of A and B by x_1 and x_2, respectively. The following are the first and the second degree (or order) polynomials in two variables:

$$y = b_0 + b_1x_1 + b_2x_2 \text{ (first order)}$$

$$y = b_0 + (b_1x_1 + b_2x_2) + (b_{11}x_1^2 + b_{12}x_1x_2 + b_{22}x_2^2) \text{ (second order)}$$

In the second order polynomial, the coefficients b_{11}, b_{12}, b_{22} could have been replaced by b_3, b_4, b_5. The double subscript, however, reminds us that these are the coefficients for the quadratic terms. Just as the second order model is obtained from the first order model by adding the second order (or quadratic) terms, we similarly obtain the third order model by adding the cubic terms $(b_{111}x_1^3 + b_{112}x_1^2x_2 + b_{122}x_1x_2^2 + b_{222}x_2^3)$ to the second order model.

In partitioning the d.f. in a 2 x 2 factorial, we are in essence fitting the model $y = b_0 + b_1x_1 + b_2x_2 + b_{12}x_1x_2$, an incomplete second order model. (With only two levels, we cannot estimate squared terms.)

In a 3 x 3 factorial, the 2 d.f. for each of the two main effects may be further partitioned into linear and quadratic terms. The 4 d.f. for the A x B interaction may be partitioned into products of the linear and quadratic terms of the main effects. Therefore, we are fitting the model

$$y = b_0 + (b_1x_1 + b_{11}x_1^2) + (b_2x_2 + b_{22}x_2^2) + (b_{12}x_1x_2 + b_{122}x_1x_2^2 + b_{112}x_1^2x_2 + b_{1122}x_1^2x_2^2),$$

(main effects of A) (main effects of B) (interaction A x B)

which is a second order model plus two cubic and one quartic terms.

Table 4 gives the analysis of variance of a randomized block experiment with b blocks and t treatments, with the t treatments forming a $p \times q$ factorial. This table should be compared with Table 1 for one quantitative factor. (If m measurements were made on each experimental unit, we will assume that they have been averaged; otherwise, there will be an extra line in the analysis of variance, as in Table 1.) The 2 d.f. for linear regression may be further partitioned to show the individual contributions from x_1 and x_2 separately.

They are partitioned similarly for quadratic and cubic regressions. The sums of squares in the s.s. column usually are called the *sequential* sums of squares. For example, ss(qr) is not the total quadratic regression s.s.; it is the *additional* s.s., after fitting a linear model. In other words, ss(qr) is the difference in regression sums of squares between fitting a linear model and a full quadratic model. If the true model (true state of nature) is linear, ms(qr), ms(cr), and ms(ℓ of) will be almost the same as ms(e), the error m.s. The quadratic model has 5 coefficients (other than the intercept b_0); therefore, it has 5 d.f. and its s.s. is obtained by adding ss(ℓ r) and ss(qr). If $p = q = 5$ (i.e., a 5×5 factorial), $t = pq = 25$ and "lack of fit" has $(t - 10) = 15$ d.f. If we are certain that a cubic model is adequate, and this is usually so, we do not need any replication. We can use ms(ℓ of) as the error m.s. in making tests of significance. With replication, however, we can test the cubic model. The extension of Table 4 to three or more quantitative factors should be obvious.

Table 4. *Analysis of variance of a randomized block experiment with 2 quantitative factors*

Sources of variation	d.f.	s.s.	m.s.
Blocks (B)	$b - 1$	ss(b)	ms(b)
Treatments (T)	$t - 1$	ss(t)	—
Linear regression	2	ss(ℓ r)	ms(ℓ r)
Quadratic reg. (additional)	3	ss(qr)	ms(qr)
Cubic reg. (additional)	4	ss(cr)	ms(cr)
Lack of fit	$t - 10$	ss(ℓ of)	ms(ℓ of)
Error (B × T)	$(b - 1)(t - 1)$	ss(e)	ms(e)
Total	$bt - 1$	ss(T)	

Since getting the various s.s. is extremely tedious on a desk calculator, a computer is necessary. If the levels of A and B are equally replicated and equally spaced (e.g., 5, 10, and 15 units for A and 100, 200, and 300 p/m for B), we can use orthogonal polynomials, as in the one-factor case. We illustrate this with a 3×3 factorial. From Section 2.3.1, we know how to obtain the linear and quadratic regression s.s. for A and for B, using either the means or the sums for the levels of A and of B. Table 5 gives the coefficients for getting the s.s. corresponding to x_1x_2, $x_1^2x_2$, $x_1x_2^2$, and $x_1^2x_2^2$. The coefficients will operate on the treatment means as usual. For example, if we denote the treatment means by $\bar{y}_1, \ldots, \bar{y}_9$ in the order shown in Table 5, the s.s. corresponding to x_1x_2 (or $A_L \times B_L$) is, from Equation (2.1) in Section 2.1, equal to $b(\bar{y}_1 - \bar{y}_3 - \bar{y}_7 + \bar{y}_9)^2/4$, where b is the number of observations in each mean. We also can use the coefficients in Table 5 to get the s.s. for A_L, A_Q, B_L and B_Q, but these can be obtained more easily from the three means for the three levels of A, and similarly for B. The reader should verify that the coefficients for the components of the main effects are similar to those given in Table 2. As before, the coefficients for interactions are the products of corresponding coefficients for the main effects. With Table 5 as an example, the reader should have no difficulty in extending this to a 3×4 or 4×5 factorial, or to more than two factors. As an exercise, the reader should write the coefficients for a $2 \times 3 \times 3$ factorial.

Table 5. *Orthogonal polynomials for 3 × 3 factorial (equally spaced)*

	Treatments								
	A=1			A=2			A=3		
B:	1	2	3	1	2	3	1	2	3
x_1 or A_L:	−1	−1	−1	0	0	0	1	1	1
x_1^2 or A_Q:	1	1	1	−2	−2	−2	1	1	1
x_2 or B_L:	−1	0	1	−1	0	1	−1	0	1
x_2^2 or B_Q:	1	−2	1	1	−2	1	1	−2	1
x_1x_2 or $A_L \times B_L$:	1	0	−1	0	0	0	−1	0	1
$x_1^2x_2$ or $A_Q \times B_L$:	−1	0	1	2	0	−2	−1	0	1
$x_1x_2^2$ or $A_L \times B_Q$:	−1	2	−1	0	0	0	1	−2	1
$x_1^2x_2^2$ or $A_Q \times B_Q$:	1	−2	1	−2	4	−2	1	−2	1

As in the one-factor case, if regression (whether linear or quadratic) is significant, then *all* treatments are different and no multiple comparison procedure is necessary. Suppose a second order model is necessary and sufficient. We can use this model for *interpolation;* i.e., to predict the response y at any point *within* the range of the values of the two factors used in the experiment. Polynomials are notoriously bad for *extrapolation.* We also can find the combination of values of x_1 and x_2 that will optimize (maximize or minimize) y. To do this, we differentiate y with respect to x_1 and x_2, set these two derivatives to zero, and solve the two resulting equations. The solution is:

$$x_1^* = (2b_1b_{22} - b_2b_{12})/(b_{12}^2 - 4b_{11}b_{22})$$

$$x_2^* = (2b_2b_{11} - b_1b_{12})/(b_{12}^2 - 4b_{11}b_{22}).$$

These values of x_1^* and x_2^* (if the true values of the b's are known) will optimize y. The *estimated* optimum value of y is obtained by putting the estimated values of x_1^* and x_2^* (in terms of the estimated b's) into the second order model.

If the two factors are two kinds of fertilizers, say, the optimum y may require such a large amount of both fertilizers that it will not be economically optimum. Instead of fitting a model to the yield y, perhaps we should fit a model to z, the yield per dollar of fertilizers applied, and optimize z.

If the response surface (value of y as x_1 and x_2 vary) is highly peaked at the optimum, we should not stray far from the optimum combination of x_1 and x_2 because y will drop sharply. On the other hand, if the response surface is rather flat near the optimum, we can depart from the optimum condition without any appreciable decrease in y and the other combinations may be more convenient. One way to study the response surface is to draw *contours*. Suppose the estimated optimum value of y is 138, say. We can set y = 135, 130, 125, etc., in the second order model. These values will give us the sets of values of x_1 and x_2 that will give an estimated yield of 135, 130, etc.

We also can use the equation to estimate the difference in the response at two different points. For example, for the same value of x_1 but different values of x_2 (x_2^* and x_2', say), the difference in the responses is $y(x_1,x_2^*) - y(x_1,x_2') = (x_2^* - x_2')\, b_2 + x_1(x_2^* - x_2')\, b_{12} + (x_2^{*2} - x_2'^2)\, b_{22}$, and its variance is $(x_2^* - x_2')^2\, V(b_2) + x_1^2(x_2^* - x_2')^2\, V(b_{12}) + (x_2^{*2} - x_2'^2)^2 V(b_{22}) + 2x_1(x_2^* - x_2')^2\, Cov(b_2,b_{12}) + 2(x_2^* - x_2')\,(x_2^{*2} - x_2'^2)\, Cov(b_2,b_{22}) + 2x_1(x_2^* - x_2')\,(x_2^{*2} - x_2'^2)\, Cov(b_{12},b_{22})$. Similarly, we can estimate $y(x_1^*,x_2) - y(x_1',x_2)$ and $y(x_1^*,x_2^*) - y(x_1',x_2')$, and their standard errors. Variances and covariances of the regression coefficients will be included in the computer printout from a good regression analysis program.

We conclude by mentioning a question of experimental design. Box and Wilson (*1951*) pointed out that the squared terms in the second order model are estimated with relatively low precision in a 3×3 factorial. Box and his coworkers have developed so-called response surface designs. The texts mentioned previously for fractional factorials also contain discussion on response surface methodology. Further references are Box and Hunter (*1958*) and Myers (*1971*).

2.4. *Mixed Factors*

Consider two factors A and B, with p and q levels respectively, where A is qualitative and B is quantitative. An example would be an experiment comparing several varieties of peanuts and several rates of a fertilizer, or destruction rates of a certain bacteria at different temperatures, using several culture media.

Table 6 shows the analysis of variance of a randomized block experiment, showing the partitioning of the d.f. for the pq treatments. We have partitioned the d.f. for the main effects of B into linear and quadratic regression only, but a higher polynomial also may be fitted. If the levels of B are spaced equally, $ss(B_L)$ and $ss(B_Q)$ will be easy to get, using orthogonal polynomials, and $ss(B_R)$ will be obtained by difference, using ss(B). If the levels of A are such that meaningful orthogonal contrasts can be formed among them (before looking at the data), we should partition its d.f. accordingly, and also the d.f. for Ax B_L, etc.

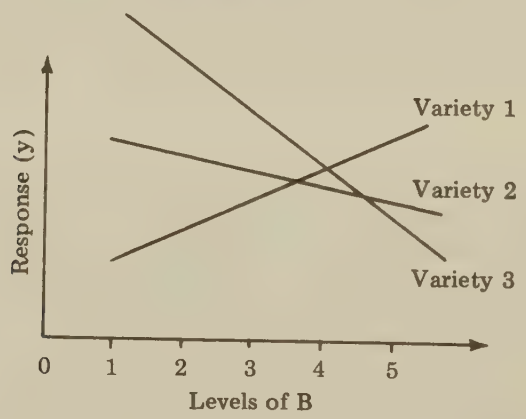
Table 6. *Analysis of variance of a randomized block experiment with 1 qualitative (A) and 1 quantitative (B) factor*

Sources of variation	d.f.	s.s.	m.s.
Blocks	b–1	ss(b)	ms(b)
Treatments (T)	pq–1	ss(t)	—
A	p–1	ss(A)	ms(A)
B	q–1	ss(B)	—
$B_{Lin.}$	1	$ss(B_L)$	$ms(B_L)$
$B_{Quad.}$	1	$ss(B_Q)$	$ms(B_Q)$
B_{Rest}	q–3	$ss(B_R)$	$ms(B_R)$
A × B	(p–1) (q–1)	ss(AB)	—
A × $B_{Lin.}$	p–1	$ss(AB_L)$	$ms(AB_L)$
A × $B_{Quad.}$	p–1	$ss(AB_Q)$	$ms(AB_Q)$
A × B_{Rest}	(p–1) (q–3)	$ss(AB_R)$	$ms(AB_R)$
Error (B × T)	(b–1) (pq–1)	ss(e)	ms(e)
Total	bpq–1	ss(T)	

If a computer is not available and the levels of B are equally spaced, one way of getting $ss(AB_L)$ is as follows. Suppose p = 4, so that A has three d.f. Arbitrarily partition these three d.f. orthogonally. A convenient set of coefficients is $A_1 = (1, -1, 0, 0)$, $A_2 = (1, 1, -2, 0)$, and $A_3 = (1, 1, 1, -3)$. Multiply these coefficients by those for B_{Linear}, to get $A_1 \times B_{Lin.}$, $A_2 \times B_{Lin.}$, and $A_3 \times B_{Lin.}$, analogously to Table 5. The sum of the sums of squares for these three interactions is $ss(AB_L)$. We get $ss(AB_Q)$ similarly. Davies (*1956*) discusses another method and gives two numerical examples.

Notice that if p = 5 and q = 6, say, A × B_{Rest} has 12 d.f. If this can be assumed not to exist, which is quite reasonable, we can use its m.s. as error and need no replication (i.e., b = 1).

As before, it does not make very much sense to test main effects of A and B if their interaction is significant. If A × B_{Linear} interaction is significant, the linear regression on the levels of B (i.e., the slope of the line) is not the same for all levels of A, as shown in the following diagram. (This is an extreme case; usually the signs of the slopes are all the same.) Similarly, if A × $B_{Quadratic}$ is significant, the curvature of the regression curve depends on the level of A. If there is no interaction, the lines (or curves) will be essentially parallel, possibly differing only in heights to reflect different effects of the qualitative factor. If it is impossible (*a priori*) to partition the d.f. for A, the levels of A may be compared using a multiple comparison procedure.



In comparing t (three or more) treatments, the null hypothesis tested is that the t true means are all equal (H_0: $\mu_1 = \mu_2 = . \ . \ . = \mu_t$). The alternative hypothesis, in general, is not that the t means are all unequal (although this may be true), but merely that they are not all equal. (For example, all but one treatment may be equal.) The next step, therefore, is to determine which treatments are different, using a so-called multiple comparison procedure. We suppose that it is impossible (*a priori*) to partition meaningfully the degrees of freedom for treatments; otherwise, the problem belongs in Chapter 2 and no multiple comparison procedure is needed.

3.1. *Error Rates*

Before describing the multiple comparison procedures that have been proposed, we will discuss the question of error rates further. We should not test all possible pairs of means with the ordinary or Student's t-test because it is relatively easy to commit a Type I error (saying two treatment means are unequal when, in fact, they are equal). For example, if we carry out a supposedly 5% t-test (i.e., 5% probability of committing a Type I error) with, say, 40 degrees of freedom for error mean square to compare all possible pairs, the probability is actually 20%, 35%, 48%, 59%, and 68% that the extremes (largest and smallest) of 4, 6, 8, 10, and 12 means, respectively, will be declared significantly different, when the true means are, in fact, all equal (David *1962*, p. 145).

When comparing three or more treatments in an experiment, there are at least two kinds of Type I error rates, based on the comparison and the experiment as the basic counting units. These rates are defined as:

Comparisonwise Type I error rate = (Number of comparisons incorrectly declared significant)/(Total number of nonsignificant comparisons tested).

Experimentwise Type I error rate = (Number of experiments with one or more comparisons incorrectly declared significant)/(Total number of experiments with at least two equal means).

If each experiment has only two treatments, these rates are identical.

Suppose Statistician A always performs his statistical tests at the 5% comparisonwise level. Each true or nonsignificant comparison will have a 5% probability of being incorrectly declared statistically significant. If, in his professional career, he makes N comparisons altogether and in M of these the null hypothesis is true, then approximately 5% of those M true comparisons will result in rejections.

In the experimentwise error rate, the experiment is the unit and no distinction is made between incorrectly rejecting one comparison and incorrectly rejecting, say, 10 comparisons in the same experiment. A Type I error is committed for the whole *experiment* if a Type I error is committed for one or more of the *comparisons* within that experiment. It does not distinguish between an experiment with true means equal to 0, 0, 0, 0, 1, say, and one with true means 0, 0, 1, 2, 3. It is easier to make one or more incorrect rejections in the former experiment, when comparing all possible pairs of means. It also does not distinguish between an experiment with 2 treatments and one with 20 treatments. Intuitively, we might feel that an incorrectly rejected comparison is more serious in an experiment with 2 treatments than in one with 20 treatments. The former experiment has only 1 comparison, while the latter has 190 possible comparisons. All other things being equal, it is obviously much easier to make one incorrect rejection in a large experiment with many treatments than in a small experiment with few treatments. Thus, a 5% experimentwise error rate is much more stringent than a 5% comparisonwise error rate. The relative frequency interpretation is as follows. If Statistician B always uses a 5% experimentwise error rate and throughout his career he analyzes N experiments and in M of these at least two treatment means are equal (so that an incorrect rejection is possible), then in approximately 5% of the M experiments, one or more comparisons will be rejected incorrectly.

For orthogonal comparisons and infinite d.f. for the error m.s. (so that the tests will be statistically independent), the experimentwise error rate (E, say) is related to the comparisonwise error rate α and the number of treatments t as follows:

$$E = 1-(1-\alpha)^{t-1}; \ \alpha = 1-(1-E)^{1/(t-1)}. \tag{3.1}$$

If $\alpha = .05$, this equation gives E = .05, .0975, .1426, .1855, .2263, .2649, .3017, .3366, .3698, .5124, and .6227 for t = 2, 3, . . . , 9, 10, 15, and 20, respectively. Thus, if we test each of the 9 orthogonal comparisons at the 5% level, in an experiment with t = 10 treatments (and the null hypothesis H_0 is true), the probability of rejecting (incorrectly) one or more comparisons is 36.98%. The overall protection against incorrectly rejecting any of the nine comparisons is 63.02% in this example.

If E = .05, the preceding equation gives α = .05, .0253, .0169, .0127, .0057, .0037, and .0028 for t = 2, 3, 4, 5, 10, 15, and 20, respectively. Thus, if we wish to hold the experimentwise error rate to 5% (i.e., 5% probability of rejecting one or more orthogonal comparisons in an experiment where all treatments are equal or, equivalently, 95% protection against incorrectly rejecting any comparison), we have to make each comparison at α = .0057 (i.e., the 0.57% level) if there are 10 treatments in the experiment.

There is no rigid rule or criterion that enables us to decide whether a comparisonwise or an experimentwise error rate is more appropriate. It is mostly a subjective choice. An experimentwise rate is more conservative in that fewer Type I errors (false significances) will be made; however, more Type II errors (failure to detect true differences) will be made. A similar problem exists in choosing the significance level α in the simple two-treatment case. Should α be taken to be .05 or .01? In situations where incorrectly rejecting one comparison may vitiate the entire experiment or incorrectly rejecting one comparison is as serious as incorrectly rejecting 10 comparisons, an experimentwise error rate is more pertinent. A comparisonwise error rate should be used if one faulty inference does not affect the remaining inferences from the same experiment. The author favors comparisonwise error rates in general. For further discussion of error rates, see Tukey (*1953b*), Harter (*1957*), and Federer (*1961*).

We shall now describe the multiple comparison procedures in turn. Some textbooks that contain a discussion of this topic are Federer (*1955*), Steel and Torrie (*1960*), Scheffé (*1959*), Seeger (*1966*), Kirk (*1968*), Bancroft (*1968*), and Miller (*1966*). Some review papers on this topic are Hartley (*1955*), Cornell (*1971*), Gill (*1973*), Games (*1971*), Ryan (*1959*), O'Neill and Wetherill (*1971*), Thomas (*1973*), Waldo (*1976*), etc. The O'Neill and Wetherill paper has a bibliography of 234 references, classified into 15 categories (multiple range tests, error rates, simultaneous confidence intervals, etc.). Thomas has an unpublished bibliography on multiple comparison techniques (available from him) containing about 300 references up to 1970.

3.2 *Fisher's Protected and Unprotected LSD Methods*

Fisher's protected LSD (least significant difference) procedure is to be applied *only* if the overall F test for treatments is significant. It consists of applying the ordinary Student's t test to any pair of means $\bar{y}_i$ and $\bar{y}_j$. Let s^2 be the error mean square (with ν degrees of freedom) from the analysis of variance table, and n_i and n_j be the number of replications of treatments i and j, respectively. The two treatments will be declared different if the two observed means y_i and y_j differ (in absolute magnitude) by more than the LSD given by

$$\text{LSD} = t(\alpha,\nu)\sqrt{s^2[(1/n_i) + (1/n_j)]}, \tag{3.2}$$

where $t(\alpha,\nu)$ is the tabulated two-sided (100 α)% value of the t-distribution with ν degrees of freedom; e.g., t(.05, 30) = 2.04.

Besides permitting unequally replicated treatments, the procedure is applicable for interval estimation. Thus, the $100(1 - \alpha)$% confidence interval for $(\mu_i - \mu_j)$ is $(\bar{y}_i - \bar{y}_j) \pm$ LSD. (Note that if the difference between $\bar{y}_i$ and $\bar{y}_j$ is less than the LSD, the confidence limits will have different signs so that the hypothesis of equal means is accepted. Recall the connection between hypothesis testing and interval estimation mentioned in chapter 1.) A third desirable feature is its ease of application, especially if all treatments are replicated equally. The LSD for *all* pairs of treatments is $t(\alpha,\nu)\sqrt{2s^2/n}$, where n is the common number of replications. (It is possible for the overall F test to be significant but none of the t tests for the pairwise differences to be significant. See Miller (*1966*, page 91).

To illustrate the method we will use the data in Duncan (*1955*) from a randomized block experiment with six blocks and seven treatments (varieties of barley). The analysis of variance gave a treatments mean square of 366.97 (with 6 d.f.), an error mean square (s^2) of 79.64 (with ν = 30 d.f.), with a highly significant F ratio of 4.61. The means (in bushels per acre) of the seven varieties, given below, have been relabeled A through G in increasing order.

49.6	58.1	61.0	61.5	67.6	71.2	71.3
A	B	C	D	E	F	G

With $\nu = 30$ and taking α to be 0.05, $t(\alpha,\nu) = 2.04$ and the LSD $= 2.04 \times \sqrt{2(79.64)/6} = 10.51$. Any two means differing by more than 10.51 will be significantly different at the 5% level. We systematically test G −A, G −B, G −C, G −D, G −E, G −F; F −A, F −B, . . . , F −E; E −A, . . . , E −D; . . . ; B −A. In practice, of course we may not need to test all possible pairs. For example, once we have found G − C = 10.3 to be less than the LSD, we need not test G −D, G −E, and G −F, for these cannot be significant. The results usually are presented by underscoring (means underscored by the same line are not significantly different) or by using superscripts (means having the same superscript are not significantly different). For the preceding example, the results are as follows:

49.6^{c}	58.1^{bc}	61.0^{ab}	61.5^{ab}	67.6^{ab}	71.2^{a}	71.3^{a}
A	B	C	D	E	F	G

Another way of presenting the results, which is typographically convenient, is to group the means as follows: (A,B), (B,C,D,E), and (C,D,E,F,G). Means in the same parentheses are not different. There were seven differences (GA, GB, FA, FB, EA, DA, and CA). An unpleasant feature of many multiple comparison procedures is the lack of "transitivity." In the preceding example, (A and B) and (B and C) were the same, but A and C were different.

This procedure is satisfactory if H_0 is true. However, suppose H_0 is false such that all means but one are equal, and this single mean is much larger (or much smaller) than the other (t −1) means. The overall F-test will be significant, and repeated t-tests applied to the (t −1) equal means will have a large probability of declaring some of these (t −1) means to be unequal. This objection is removed in the Newman-Keuls' procedure, to be discussed in Section 3.3.

In the unprotected LSD method, a preliminary F test need not be carried out at all, but the error rate for each individual comparison is reduced to α/m, where m is the total number of comparisons (preferably specified in advance) that we wish to make among the t treatments. If we restrict ourselves to orthogonal contrasts, m = (t −1); if we make all possible pairwise comparisons, m = t(t −1)/2. More generally, we can budget m different error rates $\alpha_1, \alpha_2, \ldots, \alpha_m$ for the m contrasts, where these add up to α. If it is more serious to incorrectly reject the i −th contrast than the j −th contrast, we would choose $\alpha_i < \alpha_j$. It can be shown (using the so-called Bonferroni inequality) that the experimentwise error rate E is at most α. Percentage points of the t-distribution for carrying out Fisher's unprotected LSD procedure may be found in Table A in the appendix, reproduced from Dunn (*1961*). Alternatively, Scheffé (*1959*, page 80) gives the following approximation (due to A.M. Peiser) for the upper (one-sided) α point of the t distribution with ν d.f.:

$$t_{\alpha;\nu} = z_\alpha + (4\nu)^{-1}(z_\alpha + z_\alpha^3),$$

where z_α denotes the upper α point of the standard normal distribution; e.g., $z_{.05} = 1.645$.

3.3. *Newman-Keuls' Multiple Range Test*

This method is applicable only in situations where all t treatments are equally replicated n times. As in Section 3.2, s^2 is the error mean square with ν degrees of freedom. This method does not have a prior significant F test as a prerequisite. To apply the method, we arrange the means in ascending order, but instead of comparing the difference between any two means with a constant least significant difference (as in Section 3.2), we test it against a variable yardstick

$$W_p = q(\alpha;\ p,\ \nu)\sqrt{s^2/n}, \tag{3.3}$$

where p (= 2, 3, . . . , t) is the number of means whose range (i.e., largest-smallest) we are testing, and $q(\alpha;\ p,\ \nu)$ is the (100 α)% point of q(p, ν), the distribution of the studentized range of p means and ν degrees of freedom. Values of $q(\alpha;\ p,\ \nu)$ are tabulated in Pearson and Hartley (*1966*) and Harter (*1960a*). They are

reproduced in condensed form in the Appendix (Table B), Beyer (*1968*), Miller (*1966*), Steel and Torrie (*1960*), etc.

For the numerical example in Section 3.2, t = 7, $\nu = 30$, and $\sqrt{s^2/n} = \sqrt{79.64/6} = 3.643$. For $\alpha = .05$, the values of q are:

p:	2	3	4	5	6	7
q(.05; p, 30):	2.89	3.49	3.85	4.10	4.30	4.46
$W_p = 3.643q$:	10.53	12.71	14.03	14.94	15.66	16.25

Fisher's LSD and W_2 are identical. We test G−A against $W_7 = 16.25$ since G−A is the range of 7 means. There are 2 ranges of 6 means (viz., G−B and F−A), and these are compared with $W_6 = 15.66$. Similarly, we test the three five-mean ranges G−C, F−F, E−A against $W_5 = 14.94$; G−D, F−C, E−B, D−A against $W_4 = 14.03$; G−E, F−D, E−C, D−B, C−A against $W_3 = 12.71$; and G−F, F−E, E−D, D−C, C−B, B−A against $W_2 = 10.53$. In practice, we need to perform much fewer tests than these, for once two means are judged to be not different, they are underscored by a line, and no further testing is made among means that are between the two means so underscored. We need only test G−A = 21.8 > W_7, G−B = 13.2 < W_6 (underscore), F−A = 21.6 > W_6, E−A = 18.0 > W_5, and D−A = 11.9 < W_4 (underscore). No further testing is necessary. The results are as follows:

$$A^a\ B^{ab}\ C^{ab}\ D^{ab}\ E^b\ F^b\ G^b \text{ or } (A, B, C, D) \text{ and } (B, C, D, E, F, G).$$

This method gives only 3 significant pairs (G−A, F−A, and E−A), compared to 7 pairs from the LSD method. The Newman-Keuls' procedure is intuitively more appealing than the LSD method. One feels that the difference between the extremes of 7 means should pass a more stringent test than the difference between the extremes of, say, 3 means. The method has the disadvantage of not being amenable to interval estimation. The error rate is confusing because it is neither experimentwise nor comparisonwise. At *each* stage of testing (range of t means, (t−1) means, etc.), the probability of rejecting the hypothesis of equal means, if true, is α.

3.4. *Tukey's HSD Method and Multiple Range Test*

Tukey's original HSD (honestly significant difference) procedure (*1951, 1953*) requires equal replications. It has the simplicity of Fisher's LSD method in having a constant yardstick with which to test all pairs of treatment means. The HSD is calculated as the W_p of the Newman-Keuls procedure, with p taken at its maximum value (i.e., with p = t, the total number of treatments). Thus, two treatments are declared to be different (in their effects) if the absolute magnitude of the difference between their means exceeds

$$\text{HSD} = W_t = q(\alpha;\ t,\ \nu)\sqrt{s^2/n}, \tag{3.4a}$$

where the symbols are as in Equation (3.3).

In the previous example, with t = 7 treatments, error mean square $s^2 = 79.64$ with $\nu = 30$ d.f. and n = 6 replications, the HSD = $q(\alpha; 7, 30) \times 3.643 = 4.46 \times 3.643 = 16.25$, if $\alpha = .05$. Testing the difference between every pair of means against 16.25, we get results that are identical to those given by the Newman-Keuls procedure. In general, we shall get fewer significant differences from Tukey's method. Since error rate of Tukey's HSD method is experimentwise, Hartley (*1955*) recommends that α be taken as 0.10 or higher.

Tukey's HSD procedure also can be used to construct *simultaneous* confidence intervals for *all* pairs of treatment differences as follows:

$$\text{Prob. } \{(\mu_i - \mu_j) \text{ lies within } (\bar{y}_i - \bar{y}_j) \pm \text{HSD: } i,j = 1,2, \ldots, t\} = (1-\alpha). \tag{3.4b}$$

In words, Equation (3.4b) states that the probability is 0.95 that *all* of the following statements are true:

$\mu_G - \mu_A = (71.3 - 49.6) \pm 16.25$; $\mu_G - \mu_B = (71.3 - 58.1) \pm 16.25$; . . . ;

$\mu_G - \mu_F = (71.3 - 71.2) \pm 16.25$; $\mu_F - \mu_A = (71.2 - 49.6) \pm 16.25$; . . . ;

$\mu_F - \mu_E = (71.2 - 67.6) \pm 16.25$; . . . ; $\mu_B - \mu_A = (58.1 - 49.6) \pm 16.25$.

Equation (3.4b) can be generalized to simultaneous confidence intervals for linear contrasts among the t treatment population means, as shown in Equation (3.4c).

$$\text{Prob.} \left\{ \sum_{i=1}^{t} c_i\mu_i \text{ lies within } \sum_{i=1}^{t} c_i\bar{y}_i \pm \tfrac{1}{2}(\text{HSD}) \sum_{i=1}^{t} |c_i| \right\} = (1-\alpha), \tag{3.4c}$$

for *all* sets of coefficients $(c_1, c_2, \ldots, c_t)$ satisfying $\Sigma c_i = 0$. (There is an uncountable infinity of such sets.) Equation (3.4c) immediately reduces to (3.4b) if the contrast is a pairwise difference, for then one coefficient is $+1$, another is -1, and the rest are zero. Equation (3.4c) also enables us to test a more general hypothesis H_0: $\Sigma c_i\mu_i = d$ (specified). We reject H_0 if the confidence limits for the contrast exclude d. Gabriel (*1964*) shows that at least one contrast will be significant if, and only if, the overall F test is significant. This is not true if the contrasts are restricted to paired differences only.

To overcome the conservativeness of his HSD procedure, Tukey also has proposed a multiple range test, using the average of his HSD and the Newman-Keuls statistic as the test criterion. Thus, the range of p ranked means is tested against

$$\tfrac{1}{2}[q(\alpha; p, \nu) + q(\alpha; t, \nu)]\sqrt{s^2/n}. \tag{3.4d}$$

Spjøtvoll and Stoline (*1973*) and Hochberg (*1975, 1976*) have extended Tukey's HSD procedure to allow unequal variances or unequal sample sizes. If sample sizes are unequal, two approximate procedures are to use the harmonic mean of the sample sizes (reciprocal of the arithmetic mean of the reciprocals of the sample sizes) or to replace the estimated variance of a mean (s^2/n) in Equation (3.4a) by the average of the variances of the two means concerned, viz., $s^2[(1/n_i) + (1/n_j)]/2$, as in Kramer's (*1956*) modification of Duncan's multiple range test. Keselman, Toothaker, and Shooter (*1975*) found that these two methods "have the same sensitivity for detecting real mean differences."

3.5. *Scheffé's Method*

Like Tukey's HSD, Scheffé's (*1953*) procedure is applicable to general contrasts, and not just paired comparisons. Since it employs an experimentwise error rate, Scheffé (*1959*, page 71) suggests taking $\alpha = .10$. Scheffé's procedure is more general than Tukey's in being able to handle unequal replications. Let n_i be the number of replications of the i−th treatment. The contrast $C = \sum_{i=1}^{t} c_i\mu_i$ will be estimated by $\hat{C} = \Sigma c_i\bar{y}_i$, with variance estimated by

$$\hat{V}(\hat{C}) = s^2\Sigma\,(c_i^2/n_i), \tag{3.5a}$$

where s^2 is the error mean square (from the analysis of variance table) with ν degrees of freedom, say. The 100 $(1 - \alpha)$% simultaneous confidence intervals for *all* contrasts C (uncountable infinity of them, obtainable by varying the set of coefficients $c_1, c_2, \ldots, c_t$) are

$$\hat{C} \pm \sqrt{(t-1).F(\alpha;t-1,\nu).\hat{V}(\hat{C})}, \tag{3.5b}$$

where $F(\alpha; t-1,\nu)$ is the upper $(100\,\alpha)$% point of the F-distribution with $(t-1)$ and ν degrees of freedom (for numerator and denominator, respectively). As an example, $F(.05;6, 30) = 2.42$. For pairwise differences ($\Sigma c_i^2 = 2$) and equal replications ($n_i = n$), Equation (3.5a) reduces to

$$\hat{V}(\bar{y}_i-\bar{y}_j) = 2s^2/n. \tag{3.5c}$$

From Equation (3.5b), the $100(1-\alpha)$% simultaneous confidence interval for *all* paired differences $(\mu_i-\mu_j)$ (for all i and j) is

$$(\bar{y}_i-\bar{y}_j) \pm \sqrt{(t-1).F(\alpha;t-1,\nu).\,(2s^2/n)}. \tag{3.5d}$$

Equation (3.5d) can be used to test the significance of the difference between two means μ_i and μ_j. We declare these to be different if the sample means $\bar{y}_i$ and $\bar{y}_j$ differ in absolute magnitude by an amount exceeding

$$S = \sqrt{(t-1).F(\alpha;t-1,\nu).(2s^2/n).} \tag{3.5e}$$

For t=2 treatments, S above is identical with the LSD since $\sqrt{F(\alpha;1,\nu)} = t(\alpha,\nu)$. Using the relationship between hypothesis testing and interval estimation, we can test the general null hypothesis $H_0: \Sigma c_i\mu_i = d$ (specified) by seeing whether d falls inside or outside the interval given in Equation (3.5b).

For the previous numerical example, taking $\alpha = .05$, we have $S = \sqrt{6 \times 2.42 \times 2(79.64)/6} = 19.63$. Two treatment sample means will be declared significantly different at the 5% level if their difference exceeds 19.63 in magnitude. (Note that this least significant difference is even larger than Tukey's HSD = 16.25. This is a general result. Tukey's procedure is preferred over Scheffé's for pairwise comparisons, but for general contrasts Scheffé's method gives a shorter interval.) Application of Scheffé's procedure to the previous numerical example gives the following results: (A,B,C,D,E) and (B,C,D,E,F,G). There are only two significant differences (G−A and F−A), compared to three differences from the Newman-Keuls and the Tukey procedures.

Equations (3.5a) and (3.5b) are directly applicable to situations where the sample means have unequal variances because of unequal replications, assuming that single observations are uncorrelated and have equal variances. For situations where the unequal variances of the sample means also may be caused by observations from the different treatments having unequal variances, Brown and Forsythe (*1974*) replace Equation (3.5a) by $\Sigma(c_i^2 s_i^2/n_i)$, where s_i^2 is the sample variance of the i−th treatment, and $F(\alpha; t-1,\nu)$ in Equation (3.5b) is replaced by $F(\alpha; t-1,f)$, where f is obtained using Satterthwaite's result on the d.f. of a linear combination of sample variances, as follows:

$$\frac{1}{f} = \sum_i f_i^2/(n_i-1)$$

$$f_i = (s_i^2/n_i)/\Sigma(s_i^2/n_i).$$

For another approximation, see Spjøtvoll (*1972*).

If the sample means are correlated, Equation (3.5b) will still hold but Equation (3.5a) must be modified to include the covariances of the sample means, as in Equation (3.5f).

Scheffé's method can be directly generalized to linear model situations, expressible in matrix notation as $\underline{y} = \underline{X}\underline{\beta} + \underline{\epsilon}$. This covers both multiple regression and analysis of variance models higher than just the one-way classification. The contrast $C = \Sigma c_i\beta_i$ will be estimated by $\hat{C} = \Sigma c_i b_i$, where the b_i's are the least squares estimates of the β_i's. The estimated variance of $\hat{C}$ is

$$\hat{V}(\hat{C}) = \Sigma \Sigma c_i c_j \text{ (estimated covariance of } b_i, b_j). \qquad (3.5f)$$

Most regression computer programs (e.g., the SAS package put out by North Carolina State University) include the estimated covariances of the estimated regression coefficients as part of the output. Equations (3.5b) and (3.5f) may now be used to construct simultaneous confidence intervals for linear contrasts or to make multiple comparisons among the β's.

3.6. *Duncan's Methods*

Of the several procedures that D.B. Duncan proposed between 1941 and 1975, we shall discuss only two—his most popular (multiple range test) and his most recent (Bayesian k-ratio LSD rule), which he hopes will supplant the former.

3.6.1. *Multiple Range Test*

This method assumes homoscedastic (equal variances) and uncorrelated means. It is very similar to the Newman-Keuls procedure, except that the protection level at each testing stage varies with p, the number of means whose range is being tested for significance. Duncan's rationale for decreasing the protection level as p increases is as follows. In experiments (factorial or otherwise) where the (p−1) degrees of freedom for the p treatments are partitioned into single degrees of freedom to correspond to (p−1) mutually orthogonal contrasts, the experimenter has no qualms about testing each contrast at the α level. Assuming for simplicity that the number of degrees of freedom for the error mean square is infinite (or quite large), the (p−1) F-ratios are statistically independent (almost). Therefore, the probability of rejecting one or more contrasts, if all p means are equal, is

$$\alpha_p = 1 - (1-\alpha)^{p-1}. \tag{3.6a}$$

Duncan (*1955*) modifies Newman-Keuls' multiple range test by using a variable level α_p as the significance level when testing the range of p means. As an illustration, with p = 9 equal means and $\alpha = .05$, the probability of incorrectly rejecting one or more of 8 orthogonal contrasts is $1-(.95)^8 = 1-.6634 = .3366$. This large probability of Type I error makes Duncan's multiple range test very powerful (large probability of detecting differences when they exist). Experimenters are often more interested in finding than in not finding significant differences among the treatments being tested. For this reason, Duncan's procedure received widespread acceptance among research workers, particularly in the agricultural sciences. As originally proposed, no preliminary significant overall F test is required. To overcome, somewhat, the objection of a possibly large Type I error probability, we may conservatively require a significant overall F test as a necessary condition for the application of the multiple range test.

In the Newman-Keuls procedure, the yardstick for testing the significance of the range of p means is $W_p = q(\alpha;p,\nu)\sqrt{s^2/n}$. In Duncan's procedure, the yardstick is similar, except that α is replaced by α_p, defined by Equation (3.6a), giving the following "shortest significant range" criterion:

$$R_p = q(\alpha_p;p,\nu)\sqrt{s^2/n}. \tag{3.6b}$$

Thus, no special tables are required if we have extensive tables of $q(p,\nu)$, the distribution of the studentized range of p means and ν d.f. However, the percentiles α_p are "awkward," being equal, for example, to .05, .0975, .1426, .1855, .2262, and .2649 if $\alpha = .05$ and p = 2,3,4,5,6, and 7, respectively. For this reason, Duncan (*1955*) tabulates $q(\alpha_p;p,\nu)$ for $\alpha = .05$ and .01; p = 2(1)10(2)20,50,100; and ν = 1(1)20(2)30,40,60,100, and ∞. More accurate and more extensive tables are given in Harter (*1960*), reproduced in Harter (*1970*). A condensed table of $q(\alpha_p;p,\nu)$ is given in the appendix as Table C, in Steel and Torrie (*1960*), etc.

To apply the method, we arrange the means in ascending order and test each pair against R_p, starting with the extremes. Once two means are declared to be not significantly different, we underline them and no further testing is made between means underscored by this line. Applied to the previous example with t = 7 means, $\nu = 30$ d.f., $s^2 = 79.64$, and each treatment equally replicated n = 6 times so that $\sqrt{s^2/n} = 3.643$, we have:

p:	2,	3,	4,	5,	6,	7
$q(.05_p;p,30)$:	2.89,	3.04,	3.12,	3.20,	3.25,	3.29
$R_p=3.643q$:	10.53,	11.07,	11.37,	11.66,	11.84,	11.99

The results of the test are:

A	B	C	D	E	F	G
49.6	58.1	61.0	61.5	67.6	71.2	71.3

In these results, $G - A = 21.7 > R_7$, the shortest significant range for 7 means; $G - B = 13.2 > R_6$; $G - C = 10.3 < R_5$, so we underline G through C and make no comparisons among C,D,E,F, and G. $F - A = 21.6 > R_6$; $F - B = 13.1 > R_5$, and we need not test F − C, etc.; $E - A = 18.0 > R_5$; $E - B = 9.5 < R_4$, so underline B through E; $D - A = 11.9 > R_4$; $C - A = 11.4 > R_3$; and finally $B - A = 8.5 < R_2$, so underline A and B. Thus, the method gives seven significant differences (GA, GB, FA, FB, EA, DA, CA), compared to three significant differences from Newman-Keuls' test.

One disadvantage of this procedure is that it is not amenable to simultaneous interval estimation. If we use $(y_i - y_j) \pm R_p$ as the confidence interval for $(\mu_i - \mu_j)$, some pairs of means will have confidence intervals of different widths, even though all treatments are equally replicated.

In a sense, Fisher's LSD, Newman-Keuls' MRT, and Tukey's HSD are particular cases of Duncan's MRT. If in Equation (3.6b), we put $\alpha_p = \alpha$ and p = 2, we obtain Fisher's LSD. Tukey's HSD is obtained by putting $\alpha_p = \alpha$ and p = t; and substitution of α for α_p gives the Newman-Keuls' MRT.

If the sample sizes are unequal, Bancroft (*1968*) suggests using the harmonic mean of the sample sizes (reciprocal of the arithmetic mean of the reciprocals of the sample sizes):

$$\bar{n}_h = [(n_1^{-1} + n_2^{-2} + \ldots + n_t^{-1})/t]^{-1}.$$

Kramer (*1956*) suggests replacing s^2/n (the common variance of the sample means) in Equations (3.3), (3.4a), and (3.6b) by the average of s^2/n_i and s^2/n_j, the variances of the two sample means being tested. Equation (3.6b) becomes

$$R_p = q(\alpha_p;p,\nu)\sqrt{s^2[(1/n_i) + (1/n_j)]/2}. \tag{3.6c}$$

Kramer (*1957*) extends the procedure in an obvious manner to correlated as well as heteroscedastic means, where the variance of $\bar{y}_i$ is $c_{ii}\sigma^2$, that of $\bar{y}_j$ is $c_{jj}\sigma^2$, and their covariance is $c_{ij}\sigma^2$. The coefficients c_{ii}, c_{jj}, and c_{ij} are known, but σ^2 is unknown and is estimated as usual by the error mean square with, say, ν.d.f. from the analysis of variance. (This does not handle the situation where the unequal variances of the means are due to observations from the different treatments having unequal variances. The correlation between the means may be due to an incomplete block design or a covariate being used in the analysis.) If $\bar{y}_i$ and $\bar{y}_j$ are the extremes of p ranked treatments, then we declare these treatments to be different if their difference exceeds

$$q(\alpha_p;p,\nu)\sqrt{\tfrac{1}{2}(c_{ii} - 2c_{ij} + c_{jj})s^2} \tag{3.6d}$$

in Duncan's test, and similarly for the Newman-Keuls or the Tukey tests. Note that if the means are uncorrelated, $c_{ij} = 0$, $c_{ii} = 1/n_i$, and $c_{jj} = 1/n_j$, so that Equation (3.6d) reduces to (3.6c).

Kramer's extension of the test to correlated and heteroscedastic means is approximate; and it is also conservative, in the sense that it tends to declare two means equal when they are not. Duncan (*1957*) proposes a more powerful test, which imposes a further condition for a subset of means to be declared homogeneous.

3.6.2 *Bayesian k-ratio t (LSD) Rule*

In Fisher's protected LSD method, the result of the overall F test for treatment effects is used only in a go, no-go fashion. In Duncan's Bayesian k-ratio t or k-ratio LSD rule, the observed value of the F test statistic actually is used in calculating the LSD or the critical t value for comparing two means. If the F ratio is large (indicating heterogeneous treatments), the critical t value is reduced, thereby increasing the power of the test: and if the F ratio is small (indicating homogeneous or nearly homogeneous treatments), the critical t value is increased, making it more difficult to declare two treatments to be significantly different and thus decreasing Type I error probability. Duncan (*1975*) summarizes his earlier work (1961 and 1965) and that of his former doctoral students (Ray A. Waller and Dennis O. Dixon) at The Johns Hopkins University in 1969 and 1974.

The k-ratio t test is based on an EBALEP (empirical Bayes, additive losses, exchangeable priors) approach. The sample mean $\bar{y}_i$ is, of course, a random variable, usually assumed to be normally distributed with mean μ_i and variance σ^2/n. In Bayesian statistical inference, the population means $\mu_1, \mu_2, \ldots, \mu_t$ also are regarded as random variables, with a prior distribution that usually is assumed to be normal with some mean μ_0 and variance σ_0^2. (This may well be true experimentally and not merely conceptually, if the t treatments correspond to t varieties, say, randomly selected for field testing from a larger collection of varieties.) The term "empirical Bayes" comes about through having to use the data to estimate the parameters of the conceptual superpopulation of populations. If L_i is the loss incurred when the i−th decision is erroneous, and similarly with L_j, the additive losses assumption states that the loss incurred is $L_i + L_j$, if both the i−th and the j−th decisions are incorrect. Finally, the exchangeable prior distributions assumption states *a priori* the comparisons are "equally plausible." This rules out, for example, the case where the t treatments form a p × q factorial (where *a priori* comparisons of main effects are more likely to be significant than interaction effects) or where the t treatments correspond to t levels of a quantitative factor, where we may *a priori* expect an ordering of the true treatment means $\mu_1 \leq \mu_2 \leq \ldots \leq \mu_t$. (Of course, these two cases fall under Ch. 2, and no multiple comparison technique is appropriate.)

A novel feature of the test is the use of the ratio (denoted by k) of the relative seriousness of Type I to Type II errors. By considering the case of t = 2 treatments (where no multiple comparison problem exists), the critical value in the regular Student's t test at a given α level can be made approximately equal to that in the k-ratio t test for some value of k. In round figures, the approximate correspondence between α and k is:

α :	.10, .05, .01
k :	50, 100, 500.

Therefore, Duncan recommends that k be taken to be equal to 100 or 500, where an experimenter previously used to test at the 5% or the 1% level, respectively.

Any difference d between two means or, more generally, any contrast c among the means is significantly different from zero if the ratio d/s_d or c/s_c exceeds some critical value $t(k,F,t,\nu)$, where $s_d = \sqrt{2s^2/n}$, and s^2 is the error mean square with ν degrees of freedom, n is the constant number of replications of each treatment, and F is the observed F ratio for treatments from the analysis of variance table. (The estimated variance s_c^2 of a contrast is given in Equation (3.5a).) As indicated above, the critical t value depends on the four arguments k, F, t, and ν. (Unfortunately, we have used the same letter t to denote two entirely different things—the total number of treatments in the experiment and the t test or distribution.) Its dependence on F is awkward for tabulation because of the uncountably infinite number of values that F can take, making interpolation almost inevitable in each application. There is also no easy or explicit formula for calculating the critical value. It is the solution of an extremely complicated integral equation, which appears as Equation (3.15) in Duncan (*1975*). Table D in the appendix gives the critical values for the k-ratio t test for k = 100 and 500, taken from Waller and Duncan (*1972*). For interpolating with respect to F, Waller and Duncan (*1969*) recommend linear interpolation using $a = \sqrt{1/F}$ for $F \leq 2.4$, except when $q > 100$ and $\nu > 60$; otherwise, we use $b = \sqrt{F/(F-1)}$, for $F > 2.4$, except when $q \leq 20$ and $\nu \leq 20$, where $q = t-1$. When a cannot be used, b is used, and vice versa. Interpolation with respect to q and ν should hardly ever be necessary. If needed, the recommendation is to interpolate using q and $1/\nu$. Values of a and b are included in Table D.

For large experiments (large number t of treatments and large number ν of d.f. for error), the critical values may be approximated as follows, with b already defined above:

$$\begin{aligned} t(100, F, \infty, \infty) &= 1.72\, b(\text{for } k = 100) \\ t(500, F, \infty, \infty) &= 2.23\, b(\text{for } k = 500) \end{aligned} \qquad (3.6e)$$

Duncan (*1965*) considers Equation (3.6e) to give adequate approximation if $t \geq 15$ and $\nu \geq 30$. Equation (3.6e) shows that for large F (sign of heterogeneous treatments), two means will be declared different if their studentized difference (d/s_d) exceeds only 1.72 (for k = 100, corresponding to $\alpha = .05$), while for a small F = 1.5, say, the critical value is raised to $1.72\sqrt{1.5/.5} = 2.98$, reducing the probability of Type I error.

In the numerical example we have been considering, t = 7 treatments, error mean square $s_2 = 79.64$ with $\nu = 30$ degrees of freedom, F = 4.61, and standard error of a difference $s_d = \sqrt{2s^2/n} = \sqrt{2(79.64)/6} = 5.15$. For k = 100, q = t − 1 = 6, and $\nu = 30$, Table D gives t = 2.16 for F = 4.0 (and $b = 1.155$) and t = 2.02 for F = 6.0 (and $b = 1.095$). Interpolating for F = 4.61 (and $b = \sqrt{4.61/3.61} = 1.130$), we get the critical t value as t(100, 4.61, 7, 30) = 2.02 + (2.16 − 2.02)(1.130 − 1.095)/(1.155 − 1.095) = 2.02 + .08 = 2.10. (If we had interpolated directly with respect to F, instead of the recommended $b = \sqrt{F/(F-1)}$, the calculated value of t would be 2.12. Although t = 7 is too small to be regarded as infinite, use of Equation (3.6e) gives a calculated t of $1.72\sqrt{4.61/3.61} = 1.72\ (1.13) = 1.94$.) Instead of dividing each difference by its standard error s_d and comparing it with the k-ratio t value, it will be more convenient computationally to multiply the t value by s_d to give the corresponding k-ratio LSD = 2.10 (5.15) = 10.82 for the present problem. Any two means differing by more than 10.82 will be declared different. The results are as follows, being identical to those obtained by using Fisher's LSD method.

49.6	58.1	61.0	61.5	67.6	71.2	71.3
A	B	C	D	E	F	G

The LSD's (in multiples of s_d) from the procedures for 7 treatments and 30 d.f. for error are:

	LSD/s_d
Fisher's	2.04
Newman-Keuls' MRT ($q(\alpha;p,\nu)/\sqrt{2}$)	2.04, 2.47, . . . , 3.15
Tukey's HSD	3.15
Tukey's MRT	2.60—3.15
Scheffé's	3.81
Duncan's MRT ($q(\alpha_p;p,\nu)/\sqrt{2}$)	2.04, 2.15, . . . , 2.33
Duncan's k-ratio t test (for an observed F = 4.61)	2.10

This tabulation shows that Duncan's k-ratio LSD rule is almost as powerful as Fisher's LSD, without the latter's higher Type I error probability, for if H_0 were true, the observed F would have been smaller (equal to 2.4, say) and from Table D, the critical value for t would have been 2.42. If the treatments are very heterogeneous, the k-ratio LSD rule can be more powerful than Fisher's LSD. If F = 10, for example, the critical k-ratio t value is 1.93, compared to 2.04 for Fisher's LSD rule.

The k-ratio t test is adaptable for simultaneous interval estimation. Following Fisher's, Scheffé's, and Tukey's methods, one would expect the k-ratio confidence interval for $\delta = (\mu_i - \mu_j)$ to be $(d = \bar{y}_i - \bar{y}_j) \pm$ k-ratio LSD, where the LSD $= t(k,F,t,\nu)s_d$, but this is not so. Besides the four parameters k, F, t, ν, the LSD in the interval estimation problem also depends on the observed value of $t = d/s_d$. Unfortunately, tables are not available at present. We refer the reader to Duncan (*1975*) and Dixon and Duncan (*1975*) for details. A large sample solution for the limits is as follows:

$$[\delta_L, \delta_U] = [1-(1/F)]d \pm \sqrt{1-(1/F)}s_d t(k, \infty, \infty, \infty), \tag{3.6f}$$

where t = 1.72 (for k = 100) and 2.23 (for k = 500). Note that the point estimate of $\delta = (\mu_1 - \mu_j)$ is $[1-(1/F)](\bar{y}_i - \bar{y}_j)$. Dixon and Duncan (*1975*) think that the preceding large sample approximation is adequate if $t \geq 16$, $\nu \geq 60$, and $F \geq 6$.

Another approximation that assumes only a large observed F value (with finite t and ν) is the following:

$$[\delta_L, \delta_U] = d \pm s_d t(k, \infty, t, \nu). \tag{3.6g}$$

The values of $t(k, \infty, t, \nu)$ are independent of t and are obtainable from the last row in Table D in the appendix for k = 100 and 500.

3.7. *Studentized Maximum Modulus Procedure*

All the procedures so far discussed for simultaneous interval estimation are for contrasts among the k means (or paired differences in particular). Sometimes, the experimenter may wish to construct simultaneous confidence intervals for the population means themselves. Assume that all the sample means $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_t$ are correlated equally with correlation coefficient ρ and with possibly unequal variances $d_1\sigma^2, d_2\sigma^2, \ldots, d_t\sigma^2$, where the d's are known constants. If s^2 is the usual unbiased estimate of σ^2 with ν degrees of freedom, the probability is $\gamma = (1-\alpha)$ that μ_i lies within $\bar{y}_i \pm u(t, \nu, \rho; \gamma)\sqrt{d_i}s$, for all $i = 1, 2, \ldots, t$ simultaneously, where $u(t, \nu, \rho; \gamma)$ is the two-sided $(100\gamma)\%$ point of the maximum absolute value of the t-variate Student's t distribution with ν degrees of freedom and common correlation ρ. (Constructing a $100(\gamma)^{1/k}\%$ confidence interval for μ_i independently of the others, using data from the i-th sample only, is not efficient.)

This technique can be extended to linear combinations of the means (not necessarily contrasts). The probability is $(1-\alpha)$ that $\Sigma c_i\mu_i$ lies within $\Sigma c_i\bar{y}_i \pm u(t, \nu, \rho; \gamma)\, s|\Sigma c_i\sqrt{d_i}|$ for *all* (uncountably infinite) sets of constants $(c_1, c_2, \ldots, c_t)$. Values of $u(t, \nu, \rho; \gamma)$ are given in Hahn and Hendrickson (*1971*) for $\rho = 0, .2, .4, .5$; $\gamma = .90, .95, .99$; $t = 1(1)6(2)12, 15, 20$; $\nu = 3(1)12, 15(5)30, 40, 60$. Table E in the appendix gives the values of $u(t, \nu, 0; \gamma)$. Use of Table E in cases where $\rho \neq 0$ gives conservative results. The values for $\rho \neq 0$ are smaller than corresponding ones with $\rho = 0$.

3.8 *Comparisons Against a Control*

3.8.1. *Dunnett's Method*

In experiments comparing t treatments, one of the treatments quite often is a control (check or untreated). In these experiments, we could partition the (t-1) d.f. for treatments into 1 d.f. for comparing control against the average of the other treatments and (t-2) d.f. for comparisons among the (t-1) "real" treatments. If these (t-1) other treatments are significantly different, the 1 d.f. comparison between their average and the control may not be meaningful. The experimenter may wish to compare the control with each of the other (t-1) treatments (and not with their average). Duncan's k-ratio t test is not applicable here since the exchangeable priors (or equally plausible comparisons) assumption is not satisfied. (The difference between a control and a treatment is *a priori* likely to be larger than that between two treatments.) Dunnett

(*1955*) gives a procedure for the simultaneous interval estimation or multiple comparisons of the control with each of the others, with an experimentwise error rate. A treatment and a control are declared different if their means differ by more than $t(\alpha; q, \nu)s_d$, where s_d is the standard error of a difference, $q = (t-1)$ is the number of treatments other than control. Values of $t(\alpha; q, \nu)$ are given in Dunnett (*1964*) and reproduced for both one-sided and two-sided tests in Table F of the appendix. If we are comparing insecticides, for example, and the control is a standard one, two-sided tests would be proper since we do not know *a priori* if the new insecticides would be better or worse than the standard insecticide. More extensive tables of $\sqrt{2}t(\alpha; q, \nu)$ for one-sided tests are given in Gupta and Sobel (*1957*) for up to 50 treatments.

To illustrate the method, suppose that variety A in our numerical example is a standard variety, thus calling for two-sided tests of A against each of the others. From Table F, with 30 d.f. for error and $q = 6$ other treatments besides control, the critical t value in a 5% two-sided test is $t(.05; 6, 30) = 2.72$. The standard error of a difference is $s_d = \sqrt{2s^2/n} = 5.15$. The LSD between control and each of the others is $LSD = 2.72(5.15) = 14.0$. Since the mean of A is 49.6, any variety will be different from A, if its mean is at least $49.6 + 14.0 = 63.6$. The result is that B, C, and D are not different from A, but E, F, G are better than A. The two-sided interval estimate of the difference between a standard variety and any other variety is their observed mean difference $\pm$ 14.0.

The preceding discussion assumes equal replications. If the control is replicated n_c times and the $i-th$ treatment is replicated n_i times, we define $s_d = \sqrt{s^2[(1/n_c) + (1/n_i)]}$, which reduces to the previous definition if all replications are equal. More generally, if within treatment variances are not homogeneous, we define $s_d = \sqrt{(s_c^2/n_c) + (s_i^2/n_i)}$ and use Satterthwaite's result for getting the d.f. of a linear combination of mean squares. It may suffice to calculate only two error mean squares, one for within control and the other for within other treatments. For a refinement, see Dunnett (*1964*). Dunnett's paper also gives the following optimal allocation of experimental units. If $n_1 = n_2 = \ldots = n_{t-1} = n$, say, we should take $n_c = n\sqrt{t-1}$. Bechhofer (*1969*) generalizes this result to the case where the variances are unequal but their ratios σ_i^2/σ_c^2 $(i = 1, 2, \ldots, t-1)$ are known.

Robson (*1961*) extends Dunnett's procedure to the case of a balanced incomplete block design, giving rise to correlated treatment means.

3.8.2 *Gupta and Sobel's Method*

Using the statistic in Dunnett's method, Gupta and Sobel (*1958*) give the following procedure for selecting all treatments that are as good as or better than the control or standard treatment. The procedure guarantees a probability of at least $(1-\alpha)$ that the selected subset of treatments contains all treatments that are at least as good as the control. The rule is to include in the subset all treatments whose means $\bar{y}_i$ exceed that of the control $\bar{y}_0$ by the amount

$$(\bar{y}_i - \bar{y}_0) \geq -t(\alpha; q, \nu)s_d, \tag{3.8a}$$

where $t(\alpha; q, \nu)$ is the one-sided critical value in Dunnett's test.

In using Equation (3.8a) as the criterion, we throw away treatments that are significantly worse than control. Treatments whose sample means are slightly less than those of control (so that $\bar{y}_i - \bar{y}_0$ will be slightly negative) will be included in the subset. If we use Dunnett's test as a screening procedure, we declare the $i-th$ treatment to be as good as or better than control if

$$(\bar{y}_i - \bar{y}_0) \geq +t(\alpha; q, \nu)s_d. \tag{3.8b}$$

Comparing Equations (3.8a) and (3.8b), it is obvious that Gupta and Sobel's procedure will give a larger subset of treatments. Dunnett's method retains only those treatments that have proved themselves superior to control, while Gupta and Sobel's method discards only those treatments that have proved inferior to standard treatment.

Gupta and Sobel (*1958*) also discuss other related problems—comparing variances and binomial parameters.

Sobel and Tong (*1971*) consider the optimal allocation of observations for partitioning a set of normal populations in comparison with a control.

3.8.3 *Williams' Method*

Williams (*1971*) considers the case where the t treatments are t levels or doses of some substance, with the control corresponding to zero dose. This situation was discussed in Section 2.3.1, where the recommended analysis was either to compare zero against the average of the nonzero doses and fit a regression to the q = (t−1) nonzero levels or to fit a curve through all t doses (including zero). Williams claims there are circumstances in which the experimenter may not wish to fit a curve to the t doses. He may wish, instead, to compare zero dose against each of the other doses. As an example, he cites toxicity studies in which the aim of the experiment may be to determine the lowest dose at which there is activity. (The assumption is that the response is zero up to this "lowest dose" and increases thereafter, instead of continuously increasing from zero, slowly at first and more rapidly afterwards.) Another reason for not wishing to fit a curve may be the experimenter's unwillingness to assume a particular form (logistic, etc.) for the response function. The number of levels is usually very small (3 to 5), making model fitting rather difficult.

Dunnett's procedure may be used to compare zero with the other doses, but some power is lost in not making use of the structure in the treatments. Williams assumes a nondecreasing response function so that $\mu_0 \leq \mu_1 \leq \ldots \leq \mu_q$ if the treatments $T_0, T_1, T_2, \ldots, T_q$ are in increasing order of dosages. (If, say, the third dose (i.e., second nonzero dose) is the level at which activity first becomes noticeable, we have $\mu_0 = \mu_1 < \mu_2 \leq \ldots \leq \mu_q$.) The first step in Williams' test is to estimate μ_i (i = 0, 1, . . . , q). Because of the constraints on the μ's, μ_i is not necessarily estimated by $\bar{y}_i$, the sample mean. Bartholomew (*1961*) gives the following maximum likelihood estimates of the μ's. If $\bar{y}_0 \leq \bar{y}_1 \leq \bar{y}_2 \leq \ldots \leq \bar{y}_q$, then $\mu_i = \bar{y}_i$ (i.e., μ_i is estimated by $\bar{y}_i$). Otherwise, there is at least one i for which $\bar{y}_i > \bar{y}_{i+1}$. We replace both $\bar{y}_i$ and $\bar{y}_{i+1}$ by their weighted average

$$\bar{y}_{i,i+1} = (n_i\bar{y}_i + n_{i+1}\bar{y}_{i+1})/(n_i + n_{i+1}),$$

where n_i is the number of replications of treatment or dose i. We now have only q means $\bar{y}_0, \bar{y}_1, \ldots, \bar{y}_{i-1}, \bar{y}_{i,i+1}, \bar{y}_{i+2}, \ldots, \bar{y}_q$. If these means are in nondecreasing order, we stop and estimate μ_j by $\bar{y}_j$ (for j = 0, 1, . . ., i−1, i + 1, . . . , q) and estimate both μ_i and μ_{i+1} by $\bar{y}_{i,i+1}$. Otherwise, we repeat the averaging process, giving $\bar{y}_{i,i+1}$ a weight of $(n_i + n_{i+1})$. For instance, if $\bar{y}_{i,i+1} > \bar{y}_{i+2}$, we average them to give

$$y_{i,i+1,i+2} = [(n_i + n_{i+1})\bar{y}_{i,i+1} + n_{i+2}\bar{y}_{i+2}]/(n_i + n_{i+1} + n_{i+2})$$

as the common estimate of μ_i, μ_{i+1}, and μ_{i+2}, if the sample means are now in correct ascending order.

We now have the estimated population means $\hat{\mu}_0, \hat{\mu}_1, \ldots, \hat{\mu}_q$, where some of these may be equal, from the averaging process. Assuming equal replications for all doses (including zero), we now test

$$\bar{t}_p = (\hat{\mu}_p - \bar{y}_0)/\sqrt{2s^2/n}, \tag{3.8c}$$

taking p = q, q−1, . . . , 1 in this order, stopping as soon as we get a nonsignificant result. We declare the p−th nonzero dose to be different from control if $\bar{t}_p$ above exceeds the critical value $\bar{t}(\alpha;p,\nu)$, given in Table G in the appendix. (Note that for simplicity of statistical distribution, we test $\hat{\mu}_p$ against the unadjusted sample mean $\bar{y}_0$ and not against $\hat{\mu}_0$, even if μ_0 is not estimated by $\bar{y}_0$.) Of course, we can apply the test in the following alternative way. Declare μ_p and μ_0 different if

$$(\hat{\mu}_p - \bar{y}_0) > \bar{t}(\alpha;p,\nu)s_d. \tag{3.8d}$$

Williams (*1971*) gives an example of a randomized block experiment with 8 blocks and t = 7 doses (zero and q = 6 nonzero doses), and an error mean square $s^2 = 1.16$ with $\nu = 42$ d.f. The observed means are $\bar{y}_0 = 10.4$, $\bar{y}_1 = 9.9$, $\bar{y}_2 = 10.0$, $\bar{y}_3 = 10.6$, $\bar{y}_4 = 11.4$, $\bar{y}_5 = 11.9$, and $\bar{y}_6 = 11.7$. The effect of the substance in the experiment, if anything, can only increase the mean of the response. Since $\bar{y}_0 > \bar{y}_1$, we average these to give $\bar{y}_{0,1} = (10.4 + 9.9)/2 = 10.15$, and because this average exceeds $\bar{y}_2$, we form the weighted average $\bar{y}_{0,1,2} = (2\bar{y}_{0,1} + \bar{y}_2)/3 = 10.1$. Since $\bar{y}_5$ and $\bar{y}_6$ are not in the correct ascending order, we average them to give $\bar{y}_{5,6} = 11.8$. We thus have the following estimates of the population means.

$$\hat{\mu}_0 = \hat{\mu}_1 = \hat{\mu}_2 = \bar{y}_{0,1,2} = 10.1;\ \hat{\mu}_3 = \bar{y}_3 = 10.6;\ \hat{\mu}_4 = \bar{y}_4 = 11.4;$$

$$\hat{\mu}_5 = \hat{\mu}_6 = \bar{y}_{5,6} = 11.8.$$

The standard error of a difference is $s_d = \sqrt{2(1.16)/8} = .539$. For a test at $\alpha = .05$, Table G gives the following critical values for 40 d.f.

p :	6,	5,	4,	3,	2,	1
$\bar{t}(.05;p,40)$:	1.81,	1.80,	1.80,	1.79,	1.76,	1.68
$\bar{t}(.05;p,40)s_d$:	.98,	.97,	.97,	.96,	.95,	.91

Applying equation (3.8d),

$\hat{\mu}_6 - \bar{y}_0 = 11.8 - 10.4 = 1.4 > .98$; conclude $\mu_6 > \mu_0$.
$\hat{\mu}_5 - \bar{y}_0 = 11.8 - 10.4 = 1.4 > .97$; conclude $\mu_5 > \mu_0$.
$\hat{\mu}_4 - \bar{y}_0 = 11.4 - 10.4 = 1.0 > .97$; conclude $\mu_4 > \mu_0$.
$\hat{\mu}_3 - \bar{y}_0 = 10.6 - 10.4 = 0.2 < .96$; conclude $\mu_3 = \mu_2 = \mu_1 = \mu_0$.

The conclusion is that the fourth nonzero dose was the lowest dose at which response was observed.

Williams (*1972*) extends the procedure to handle the case where the zero dose has a different (larger) number of replications than that of the nonzero levels, for both one-sided and two-sided tests.

In general, we would recommend the regression approach of Section 2.3.1. Suppose we have the following results:

Dose:	0	1	2	3	4	5
Response:	5	7	10	15	25	40

Using the present procedure, we may conclude that treatment is first effective at dose 3. The author would rather believe that the response is increasing continuously from dose 0, gradually at first and more rapidly at higher doses. We might fit a curve and estimate the lowest dose at which the response will be at least y^*, say. If higher doses are more expensive and cost is a consideration, we could adjust the response to a per dollar basis and estimate the dose that will produce the highest adjusted response.

3.8.4 *Sequential Methods*

See Dudewicz, Ramberg, and Chen (*1975*) for a two-stage procedure when variances are unequal and unknown, and Paulson (*1962*) for a sequential procedure, assuming equal variances. In the latter, inferior treatments are dropped at each stage.

3.9. *Miscellaneous Methods*

In this section we shall discuss briefly various related techniques or merely cite their references.

3.9.1 *Bonferroni Procedure for Preselected Contrasts*

Tukey's and Scheffé's methods enable us to construct confidence intervals for an infinite number of linear contrasts among the t means so that the probability is $(1 - \alpha)$ that they are all simultaneously true. Usually an experimenter is only interested in a rather small subset of m contrasts, say. If these m contrasts are preselected and not suggested by the data, Dunn (*1961*) recommends the usual method based on the Student's t distribution to construct an interval for each contrast independently, with confidence coefficient $1 - (\alpha/m)$, so that from Bonferroni's inequality, the overall or simultaneous confidence level for all m contrasts is at least $(1 - \alpha)$, as in Fisher's unprotected LSD. Two-sided $(100\ \alpha/m)\%$ points of the t distribution are given in the paper and reproduced in Table A in the appendix. In the notation of Section 3.5, the confidence interval for each contrast is

$$\hat{C} \pm t(\alpha/m;\nu)\sqrt{\hat{V}(\hat{C})}, \tag{3.9a}$$

where $t(\alpha/m;\nu)$ is the *two-sided* $(100\ \alpha/m)\%$ point of the t distribution with ν degrees of freedom. These intervals often will be narrower than those given by Tukey's or Scheffé's methods. See also Schafer and MacReady (*1975*).

3.9.2 *Gabriel's Simultaneous Test Procedure (STP)*

Gabriel (*1964*, *1969a*) gives a procedure for testing the homogeneity of the $(2^t - t - 1)$ subsets (with at least two means) from a set of t means. Let P be any subset containing at least two treatments and S_p^2 be the treatment sum of squares for those treatments in P. These treatments will be declared to be different if

$$S_p^2 > (t-1)s^2F\ (\alpha;t-1,\nu), \tag{3.9b}$$

where s^2 is the error mean square with ν d.f. from the analysis of variance of the complete data (with t treatments), and $F(\alpha;t-1,\nu)$ is the upper $(100\ \alpha)\%$ point of the F distribution with $(t-1)$ and ν d.f. Note that the critical value of F in Equation (3.9b) is that for the complete data so that the righthand side is identical for all subsets.

The error rate is experimentwise. If H_0 is true (all t means are equal), the probability is only α that one or more of the $(2^t - t - 1)$ subsets will be declared incorrectly to be heterogeneous. The procedure also has the following nice property. Any set containing a significant subset is itself significant. (However, the converse is not necessarily true, and it is possible for a significant set to contain no significant proper subsets.) Because of this property, it is not necessary to test all subsets. For example, if the set (A, B, C) is significant, the set (A, B, C, D) will be significant; and if (E, F, G) is not significant, the subsets (E, F), (E, G), and (F, G) also will be not significant.

The 1964 paper has a numerical example. Tukey's HSD method, which is conservative compared with Newman-Keuls' or Duncan's multiple range tests, found two significant pairs. Gabriel's STP and Scheffé's test found all subsets of two means (i.e., all paired differences) to be not significant. Generally, a set P will be declared significant by Gabriel's STP if and only if some contrast involving only those means in P is judged significant by Scheffé's procedure.

3.9.3 *Kurtz-Link-Tukey-Wallace Range Procedure*

The analysis of variance is based on sums of squares. For computational convenience, analogous procedures based on ranges are available. Kurtz, Link, Tukey, and Wallace (*1965*) give a similar shortcut procedure for multiple comparisons. This paper also has an interesting general discussion on the philosophy of multiple comparisons.

3.9.4 *Covariance Adjusted Means*

For multiple comparisons of adjusted treatment means in an analysis of covariance, see Kramer (*1957*), Halperin and Greenhouse (*1958*); Scheffé (*1959, pp. 209–213*); Bancroft (*1968, Section 8.7*); and Thigpen and Paulson (*1974*).

3.9.5 *Procedures for Two-Way Interactions*

Suppose that the t treatments are in the form of a $p \times q$ factorial, both factors being *qualitative*. The partitioning of the pq-1 degrees of freedom for the $t = pq$ treatments is discussed in Section 2.2. Harter (*1970*) gives a procedure for comparing interaction effects of the form

$$\begin{aligned} A_iB_u + A_jB_v - A_iB_v - A_jB_u &= [(A_i - A_j)B_u] - [(A_i - A_j)B_v] \\ &= [A_i(B_u - B_v)] - [A_j(B_u - B_v)], \end{aligned}$$

where A_iB_u, for example, is the mean for the i–th level of factor A and the u–th level of factor B. The preceding interaction is the difference between two differences; viz., (difference between the i–th and the j–th levels of factor A, both at the u–th level of B) minus (difference between the i–th and the j–th levels of A, both at the v–th level of B). As the second form of the expression shows, the interaction also can be written as the difference between the u–th and the v–th levels of B at the i–th level of A minus the same difference at the j–th level of A. See also Dunn and Massey (*1965*), Sen (*1969*), Johnson (*1976*), and Bradu and Gabriel (*1974*). The last paper describes three methods for testing and simultaneous interval estimation.

3.9.6 *Nonparametric Methods*

In all the methods considered so far, we have assumed that the data are distributed normally. If we cannot or do not wish to make this assumption, we must resort to nonparametric methods for separating the means. See Steel (*1959, 1961*); Dunn (*1964*); Miller (*1966, ch 4*); Rhyne and Steel (*1965, 1967*); McDonald and Thompson (*1967*); Tobach et al. (*1967*); Rizvi, Sobel, and Woodworth (*1968*); Sen (*1969*); Puri and Puri (*1969*); Slivka (*1970*); and Hollander and Wolfe (*1973*, Sections 6.3, 7.3, and 7.7).

3.9.7 *Gupta's Random Subset Selection Procedure*

In experiments where the scientist is looking for the best treatment (e.g., a plant breeder selecting a new variety for highest yield or resistance to some disease), multiple comparison techniques are inappropriate. We cited Gupta and Sobel (*1958*) in Section 3.8.2 for a method for selecting treatments that are as good as or better than a control or standard treatment. Some selected references on problems of selecting the best out of t treatments are Paulson (*1964*); Gupta (*1965*); Robbins, Sobel, and Starr (*1968*); Bechhofer, Kiefer, and Sobel (*1968*); Sobel (*1969*); Tong (*1970*); Rizvi (*1971*); Chiu (*1974a*, *1974b*); a review paper with 71 references by Weatherill and Ofosu (*1974*); Wackerly (*1975*); Santner (*1975*); and Gupta and Panchapakesan (*1971*).

Selection problems may be posed in several ways, of which the following two are the most common.

(a) Given $\delta^* > 0$ and $P^* < 1$, find a procedure that will, with probability of at least P^*, choose the population with the largest mean if this mean exceeds the second largest mean by at least δ^*.

(b) Given $1/t < P^* < 1$, find the smallest subset of the t treatments such that the probability is at least P^* that the subset will contain the best population.

The preceding formulations are referred to as the "indifference zone" and the "random subset" approaches, respectively. In (a), we are indifferent to all differences that are less than δ^*; and in (b), the number of treatments that are included in the subset is a random variable. Decision theoretic approaches (minimax, Bayesian, etc.) are also possible.

Gupta (*1965*) gives the following random subset solution. Include the i−th treatment in the subset if its sample mean $\bar{y}_i$ satisfies the condition

$$\bar{y}_i \geq \bar{y}_{max.} - t(\alpha;t,\nu)s_d, \tag{3.9c}$$

where $t(\alpha;t,\nu)$ is the one-sided critical value of Dunnett's test statistic (Section 3.8.1). Values of $t(\alpha;t,\nu)$ are given in Table F1 in the appendix, with $t = (q+1)$; e.g., if $t = 7$, we look under $q = (t-1) = 6$.

In our numerical example, we have $t = 7$, $\nu = 30$ d.f., $s_d = \sqrt{2s^2/n} = \sqrt{2(79.64)/6} = 5.15$, and $\bar{y}_{max.} = 71.3$. Taking $\alpha = 1 - P^* = .05$, the value of t(.05; 7, 30) from Table F1 with $t = 7$ (or $q = 6$) is 2.40. From Equation (3.9c), we include in the subset all treatments whose means exceed $71.3 - (2.40)(5.15) = 71.3 - 12.36 = 58.94$. Thus, we are 95% confident that the set (C, D, E, F, G) will contain the best treatment (variety).

3.9.8 *Scott and Knott's Cluster Analysis Method*

If a scientist has collected a mass of data (usually multivariate), he may wish to know if these came from one or more populations. If the latter, he would like to know into how many groups or clusters the data should be divided, and the best way of forming these groups. (For a recent paper and book on cluster analysis, see Kuiper and Fisher (*1975*) and Hartigan (*1975*).) With univariate data, we can arrange the observations in ascending order. If the data are 10, 11, 55, 56, 59, for example, they can be divided into two clusters in an obvious manner, namely (10, 11) and (55, 56, 59). In less clearcut situations, an objective criterion for grouping is required. If we know that the data came from two populations only, we can form the two groups by maximizing the sum of squares between the two groups (or equivalently, such that the sum of the within groups sums of squares is a minimum). With t observations (or means), we need only consider the $(t-1)$ possible partitions formed by dividing between two successive ordered means. The multiple range tests we have considered do, in fact, group the means, but they allow a particular mean to be in more than one group. Duncan's test, for example, groups the means in the example into (A, B), (B, C, D, E), and (C, D, E, F, G). Tukey (*1949*) was the first to consider forming nonoverlapping clusters by looking at the gaps in the ordered means and testing their statistical significance, but he retracted this procedure in his 1953 manuscript (circulated privately) on the problem of multiple comparisons.

Scott and Knott (*1974*) propose the following sequential partitioning and testing procedure. Arrange the $t = 7$ means in ascending order, denoted by A, B, C, D, E, F and G, respectively. Partition these into two groups, using the above criterion. Suppose this results in (A, B, C, D) and (E, F, G) as the two groups. Now test the null hypothesis H_0: $\mu_1 = \mu_2 = \ldots = \mu_7$ against the alternative hypothesis H_a: $\mu_i = m_1$ or m_2. (Presumably, the overall F test with 6 and ν d.f. need not be performed. The usual F statistic tests H_0 against the most general alternative that not all means are equal. The proposed procedure tests H_0 against the much more specific alternative that all the means are either m_1 or m_2, with at least one mean in each group, and, therefore, should be more powerful than the usual F test.) If H_0 is rejected, we partition (A, B, C, D) into two

groups and test the equality of these groups. The procedure is similar for (E, F, G). It is repeated until H_0 is accepted.

The test is as follows. We assume that the t means $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_t$ are uncorrelated and homoscedastic, which implies equal replications n, say. As usual, let s^2 be the estimate (with ν d.f.) of the common variance σ^2 of single observations. (In the completely randomized design, $\nu = t(n-1)$.) Suppose that the partitioning criterion forms two groups with t_1 and $t_2 = (t - t_1)$ means. The groups G_1 and G_2 will contain nt_1 and nt_2 original observations, respectively. Let T_1 be the sum of the nt_1 observations in G_1, and similarly for T_2. In the usual analysis of variance computations, the between groups sum of squares is

$$B_0 = [T_1^2/(nt_1)] + [T_2^2/(nt_2)] - [T^2/(nt)], \tag{3.9d}$$

where $T = (T_1 + T_2)$. Under the null hypothesis, the maximum likelihood estimate of σ^2 is

$$\hat{\sigma}_0^2 = [n \sum_{i=1}^{t} (\bar{y}_i - \bar{\bar{y}})^2 + \nu s^2]/(t + \nu), \tag{3.9e}$$

where $\bar{\bar{y}} = (\bar{y}_1 + \ldots + \bar{y}_t)/t$.

The test statistic is

$$\lambda = \pi(B_0/\hat{\sigma}_0^2)/[2(\pi - 2)] = 1.376\,(B_0/\hat{\sigma}_0^2). \tag{3.9f}$$

The 95% points for the distribution of λ were obtained by simulation and were found to be approximated adequately, for practical purposes, by the chi-square distribution with $\nu_0 = t/(\pi - 2) = t/(1.1416)$ d.f.

(The simulation also included the case with $\nu = 0$, for which the 95% points of λ were estimated to be 2.75, 6.60, 12.11, and 21.74 for $t = 2$, 5, 10, and 20, respectively. This shows that we can test the homogeneity of t means, even when each mean is based on $n = 1$ replication. This is, of course, impossible with the usual F test and its general alternative hypothesis since the error mean square has zero d.f. As mentioned earlier, the present λ test makes an extra assumption about the alternative hypothesis.)

In our numerical example, $t = 7$, $n = 6$, $s^2 = 79.64$ with $\nu = 30$ d.f. (design being that of a randomized block experiment). The means in ascending order were 49.6(A), 58.1(B), 61.0(C), 61.5(D), 67.6(E), 71.2(F), and 71.3(G). To find the partition with the largest between groups sum of squares, we should try, theoretically, the $t - 1 = 6$ possible partitions: (A, BCDEFG), (AB, CDEFG), (ABC, DEFG), (ABCDE, FG), (ABCDEF,G). In practice, we need try two or three possibilities only. (With a computer it is easy enough to try all $(t-1)$ partitions.) In this example, (A, BCDEFG) and (ABCD, EFG) are the two most serious candidates. It can be shown that (ABCD, EFG) is the optimum partition. Here, $t_1 = 4$, $t_2 = 3$, $T_1 = 6\,(49.6 + 58.1 + 61.0 + 61.5) = 3181.2$, $T_2 = 1260.6$, $T = T_1 + T_2 = 2641.8$, $\bar{\bar{y}} = (49.6 + \ldots + 71.3)/7 = 62.9$, and $\Sigma(\bar{y}_i - \bar{\bar{y}})^2 = 370.04$.

From Equations (3.9d) and (3.9e), $B_0 = (1381.2)^2/24 + (1260.6)^2/18 - (2641.8)^2/42 = 1602.86$ and $\hat{\sigma}_0^2 = [6(370.04) + 30(79.64)]/(7 + 30) = 124.58$. From Equation (3.9f), the test statistic is $\lambda = 1.376\,(1602.86/124.58) = 17.70$. Using the chi-square approximation with $\nu_0 = t/1.1416 = 7/1.1416 = 6.1$ d.f., the value 17.70 is significant. (The 95% point of the chi-square distribution is 12.6 for 6 d.f. and 14.1 for 7 d.f.)

We next have to partition (ABCD) and (EFG). In partitioning (EFG), t is now equal to three. For $t = 3$ means, the optimum partition is at the larger of the two gaps, giving (E, FG) with $t_1 = 1$, $t_2 = 2$, $T_1 = 405.6$, $T_2 = 855.0$, $\Sigma(\bar{y}_i - \bar{\bar{y}})^2 = 8.8866$, giving $\hat{\sigma}_0^2 = [6\,(8.8866) + 30\,(79.64)]/33 = 74.02$, $B_0 = 405.6^2/6 + 855^2/12 - 1260.6^2/18 = 53.29$, and $\lambda = (1.376)\,(53.29)/74.02 = 0.99$, which is not significant. The significance of the partition of (ABCD) into (A, BCD) is borderline. If we accept this as being significant, the final groupings are A, BCD, and EFG, which is what inspection of the means would suggest.

For another cluster analysis approach to multiple comparisons, see Jolliffe (*1975*).

3.9.9 *Multivariate Populations*

We have so far considered univariate populations only. Quite often, we may collect several kinds of measurements from each experimental unit. For example, in comparing t brands of chocolate cake mixes, we may evaluate the resulting cakes with respect to each of p characteristics (flavor, aroma, texture, moistness, etc.). As another example, we may compare t treatments (storage conditions) for degreening lemons and take

color measurements on each of p dates. We may (and sometimes do) carry out p separate univariate analyses of variance, one for each of the p characteristics or dates, but we sacrifice some power in not making use of the correlations among the p characteristics. There is also a problem with the overall significance level in making p separate analyses. Preferably, we should perform one multivariate (p-dimensional) analysis of variance. If the null hypothesis of equal mean vectors (each population mean is now a set of p numbers) is rejected, we now have two different kinds of multiple comparison problems. With respect to which of the p characteristics do the populations differ? (In the preceding cake example, do the cakes differ in flavor only, in flavor and texture only, or in all p characteristics?) We do not, of course, have this problem in univariate (p = 1) situations. We have been considering the other kind of multiple comparisons in this report (viz., which populations differ from which). These comparisons are discussed in Kramer (*1972*, Section 5.11), Gabriel (*1968*, *1969b*), Krishnaiah (*1969*), Miller (*1966*, Chapter 5), and Morrison (*1967*, Section 5.4).

3.9.10 *Subset Selection Approach to Multiple Comparisons*

We mentioned in the last paragraph of Chapter 1 that hypothesis testing is usually almost totally irrelevant. Two treatments will be declared significantly different if they are sufficiently replicated. If two means are declared significantly different, many experimenters often are misled into thinking that the difference is of *practical* importance. Reading (*1975*) applies the indifference zone formulation of subset selection problems to multiple comparisons. The experimenter specifies three quantities: P(probability that all decisions concerning pairwise means are correct, an experimentwise probability), δ_* (largest amount that two populations can differ and still be considered practically the same), and δ^* (smallest amount by which two population means must differ to be considered definitely different). The interval (δ_*, δ^*) is the indifference zone. If two treatments differ by an amount in this zone, the experimenter does not care whether the treatments are declared different or the same. Given these three quantities, Reading gives tables for the necessary sample size and the critical value that must be exceeded for the difference between two means to be declared significant. Unfortunately, at present, the tables go up to t = 4 treatments only and assume that σ^2 is known.

3.9.11 *Other Parameters and Populations*

In this publication, we have been comparing, estimating, or selecting normal populations with respect to their means. We conclude this chapter by citing selected references to similar work for other parameters and other populations.

(a) Variances of normal populations. See David (*1956*), Ryan (*1960*), Bechhofer (*1968*), and Levy (*1975a*, *1975b*) for multiple comparisons; Jensen and Jones (*1969*) for simultaneous interval estimation; Gupta (*1965*), Ofosu (*1975*), and Arvesen and McCabe (*1975*) for subset selection.

(b) Various kinds of simultaneous prediction intervals. Hahn (*1970*, *1972*).

(c) Regression coefficients. Duncan (*1970*) for multiple comparisons, and Hahn and Hendrickson (*1971*) for simultaneous interval estimation.

(d) Subset selection for normal population with the largest (or smallest) α − quantile. Barlow and Gupta (*1969*).

(e) Subset selection for normal population with the largest exceedance probability. Kappenman (*1972*) gives a method for selecting the normal population with the highest $h_i = P(X_i > c)$, where $X_i \sim N(\mu_i, \sigma_i)$ and c is a given constant.

(f) Comparison of several independent treatment mean squares against a common error mean square. See Nair (*1948*); Hartley (*1955*); and David (*1962*, pages 155–156).

(g) Subset selection for gamma populations. Gupta (*1963*).

(h) Ranking and selection of binomial populations. Gupta and Sobel (*1960*), Ryan (*1960*), Taylor and David (*1962*), Paulson (*1967*), Bland and Bratcher (*1968*), Hoel and Sobel (*1972*), and Leonard (*1972*).

(i) Multinomial populations. Goodman (*1965*) and Fienberg and Holland (*1973*) for simultaneous estimation; Bechhofer, Elmaghraby, and Morse (*1959*) for selection; and Gabriel (*1966*) for multiple comparisons.

(j) Subset selection for Poisson, negative binomial, and Fisher's logarithmic distributions. Gupta and Panchapakesan (*1971*).

(k) Multiple comparisons of regression functions. Spjøtvoll (*1972*).
(l) Multiple comparisons of logistic curves. Reiersøl (*1961*).
(m) Selection of best treatment in paired-comparison experiments. Trawinski and David (*1963*).
(n) Ranking of main effects in analysis of variance, variances of normal populations, and correlation coefficients of bivariate normal distributions. Eaton (*1967*).
(o) Interval estimation of a ranked parameter. Alam and Saxena (*1974*).
(p) Simultaneous interval estimation of contrasts among means of a multivariate normal population. Bhargava and Srivastava (*1973*).
(q) Applications to multiple regression problems. Miller (*1966*), Morrison (*1967*, Section 3.6), Wynn and Bloomfield (*1971*), Hochberg and Quade (*1975*), and Tarone (*1976*).

CHAPTER 4. *CONCLUSION*

The findings from some Monte Carlo sampling studies that have been conducted to evaluate the relative performances of the various multiple comparison procedures are summarized in this chapter. Here, we assume that multiple comparisons are appropriate, ruling out situations covered in Chapter 2, where the proper statistical technique is the partitioning of the degrees of freedom for treatments into orthogonal contrasts. When it is not possible *a priori* to form meaningful orthogonal contrasts, it is assumed that the problem is really one of multiple comparisons and not of ranking and subset selection. A plant breeder who is interested in selecting a new variety should not be concerned with multiple comparisons of all possible pairs of varieties.

Scheffé's method is the most versatile. It allows unequal replications, correlated means from covariance adjustment, general contrasts (and not just paired comparisons), and simultaneous interval estimation. The penalty for this generality is reduced power (failure to detect true differences in testing and wide confidence intervals in interval estimation of differences between two means). Tukey's HSD method also can handle general contrasts and interval estimation, but it requires equal replications and uncorrelated means. Duncan's and Newman-Keuls' multiple range tests are exact only for paired comparisons of uncorrelated means with equal replications and are not adaptable for interval estimation. The LSD easily can handle unequal replications, can be used for interval estimation, and can be extended in a simple and obvious manner to general contrasts. Duncan's Bayesian k-ratio rule is too new to have found widespread acceptance by experimental scientists. Duncan is very enthusiastic about this procedure and, in a private communication, expressed the hope that his *Biometrics* 1975 paper "will mark the beginning of the end of all of the earlier (pre-1960) α-level multiple comparison procedures."

We refer the reader to Section 3.6.2, where we tabulate the LSD's for the various procedures (in multiples of the standard error of the difference between two means). In ascending order, we have Fisher's LSD, Duncan's k-ratio rule, Duncan's MRT, Newman-Keuls' MRT, Tukey's MRT, Tukey's HSD, and Scheffé's method. (Duncan's k-ratio rule is data dependent. It may be more "reckless" than Fisher's LSD or more conservative than Tukey's HSD, depending on the observed value of the F ratio for treatments.) The above order is, therefore, in decreasing order of the number of paried comparisons that will be declared significant. If the objective is to find as many significantly different pairs as possible, Fisher's LSD is best. The problem, however, is not this simple.

There are two main difficulties in assessing the relative merits of the multiple comparison procedures. "In testing a hypothesis involving a simple two-decision situation, such as that to which the Neyman-Pearson theory is directly applicable, one compares two competing test criteria by fixing the Type I errors to be the same for both and compare the two power curves. Unfortunately, multiple-comparison procedures do not pertain to a single simple two-decision situation, but are special cases of *multiple*-decision procedures. At present there is no generally acceptable analytical method of comparing, in a manner similar to that for the two-decision situation, two competing multiple-decision test criteria." (Bancroft *1968, p. 105.*)

Another difficulty is due to the different error rates used. Tukey's and Scheffé's methods use an experimentwise error rate, while Fisher's LSD adopts a comparisonwise error rate. The multiple range tests of Duncan and of Newman-Keuls use different error rates, both of which are neither experimentwise nor comparisonwise. Duncan's k-ratio rule does not even use the concept of error rate; it uses the ratio of the relative seriousness of the two types of errors.

Because of these difficulties, the procedures have been compared using Monte Carlo sampling methods only. There is a difficulty with such empirical sampling studies. It is easy to study the probability of Type I error (declaring two equal means to be unequal) because there is, of course, only one way in which t means can be equal. It is much more difficult to compare the probability of Type II error (declaring two unequal means to be equal), because t means can be unequal in many ways. They can be all unequal (equally spaced, clustered in two or more groups, etc.), all equal but one, etc. It is unlikely that one method will be best for *all* patterns of inequality.

Balaam (*1963*) was the first to publish results of a sampling study. He considered only four means, each with five observations, in eighteen configurations: (0,0,0,0), (1,0,0,0), . . ., (6,0,0,0); (1,1,0,0), (2,1,0,0), . . ., (5,1,0,0); (2,2,0,0), (3,2,0,0), (4,2,0,0); (3,3,0,0), (4,2,1,0), and (4,4,1,0). Three procedures (LSD, Newman-Keuls', and Duncan's MRT) were compared, each with and without a significant preliminary F test. The Newman-Keuls' procedure was found inferior. The LSD was superior to Duncan's MRT, in both protected and unprotected cases, but the difference in performance was small in the protected case.

Boardman and Moffitt (*1971*) compared five procedures (LSD, Scheffé's, Tukey's HSD, Newman-Keuls' MRT, and Duncan's MRT) for testing all possible pairs of means with respect to their Type I comparisonwise and experimentwise error rates. They carried out 30 sets of 10,000 sampling experiments with $t = 2, 3, \ldots, 11$ normal populations; samples of equal sizes $n = 5$, 10, and 15; and $\alpha = .05$.

For $t = 10$ treatments, and taking $\alpha = 5\%$, the Type I comparisonwise error rate for Duncan's MRT is about 2.5%, .21% for Tukey's, and .01% for Scheffé's procedure, showing the conservativeness of the latter two procedures.

On an experimentwise basis, the error rate in Tukey's HSD and Newman-Keuls' multiple range test remains constant at 5% as t increases from 2 to 10, while for Duncan's MRT and Fisher's LSD, it increases to 38% and 63% respectively. For Scheffé's procedure, it decreases from 5% to .23%, showing conservativeness of the Scheffé procedure for pairwise contrasts. Thus, with $t = 10$ populations with equal means (and $(10 \times 0)/2 = 45$ possible pairwise comparisons), there is a 38% probability that one or more of the 45 comparisons will be declared significantly different by Duncan's procedure.

In view of this rather high experimentwise probability, Gill (*1973*) recommends that Duncan's procedure be discontinued. Of course, Gill has even stronger feelings against the LSD procedure. In defense of these two procedures, the comparison, rather than the experiment, is the basic unit for the comparisonwise adherents. One wrong conclusion will not affect the usefulness of the remaining 44 comparisons. On the other hand, the rationale of the experimentwise error rate philosophy is that *one* wrong comparison vitiates *all* of the remaining 44 comparisons. Thus, making one wrong conclusion is as serious as making 45 wrong judgments in the same experiment (is this reasonable, in most cases?). We have to ensure that *all* 45 comparisons are correct, not without having to pay a high premium, of course. For example, in a cubic lattice design with $t = 729$ varieties, (Cochran and Cox *1957*, page 423), it will be virtually impossible to ensure that *all* $(729 \times 728)/2 = 265{,}356$ paired comparisons will be judged correctly.

Because of the independence of the validity of the individual comparisons (in the comparisonwise school), we can "afford" one wrong comparison out of 45. After all, in a 5% test, there is a one in 20 chance of an incorrect rejection so that out of 45 comparisons we should expect and tolerate about two wrong conclusions. In addition to the probability of one or more wrong rejections out of 45, it will be interesting to know also the probability of two or more wrong rejections. If the probability of two or more incorrect conclusions is considerably lower than that of one or more wrong conclusions, this should remove much of Gill's objections to Duncan's MRT and Fisher's LSD procedures.

In agricultural experiments, the treatment means are much more likely to be unequal so that Type II error consideration should be at least as important as Type I error consideration. In the Boardman and Moffitt study, the procedures were applied without a prior significant overall F test, which is, in fact, a prerequisite of the Fisher's protected LSD method. Although not required for the Duncan procedure, it may be desirable to apply the procedure only after a significant F test. As Dunnett (*1970*) points out, multiple comparison procedures are techniques for ferreting out differences among the t means, and there is no reason for doing so, unless there is an indication that differences exist, either *a priori* or as evidenced by a significant F test. The experimentwise error rates for the protected Fisher's LSD and the "protected" Duncan's MRT will, of course, be 5%. See Bernhardson (*1975*).

Based on the Boardman-Moffitt study (who considered only the null case of equal means), Gill recommended Tukey's HSD and, to a lesser extent, the Newman-Keuls' procedure. In another simulation study, Carmer and Swanson (*1973*) recommended just the opposite. Their conclusions were:

> . . . that Scheffé's test, Tukey's test, and the Student-Newman-Keuls' test are less appropriate than either the least significant difference with the restriction that the analysis of variance F value be significant at $\alpha = .05$, two Bayesian modifications of the least significant difference or Duncan's multiple range test. Because of its ease of application, many researchers may prefer the restricted least significant difference.

Carmer and Swanson conducted 88,000 simulations in all, with various numbers of treatments and replicates, and different patterns of heterogeneity among the treatment means. The study "was prompted mainly by the authors' own uncertainty as to the most appropriate procedure to recommend to students and researchers in the agricultural sciences." In an earlier publication, Carmer and Swanson (*1971*) reported on 5 of the present 10 procedures.

The following multiple comparison procedures were studied:

1. LSD (unprotected)
2. TSD (Tukey's HSD)
3. SNK (Student-Newman-Keuls)
4. MRT (Duncan's multiple range test)
5. SSD (Scheffé's procedure)
6. FSD1 (Fisher's protected LSD, with the preliminary F test applied at the 1% level)
7. FSD2 (as in FSD1 but F test at 5% level)
8. FSD3 (as in FSD1 but F test at 10% level)
9. BSD (Duncan's approximate Bayesian k-ratio LSD rule for $t \geq 15$ treatment and error d.f. $\nu \geq 30$; see Equation (3.6e) of present report)
10. BET (Waller-Duncan's exact Bayesian k-ratio LSD rule)

We quote from Section 7 ("Concluding Remarks") of Carmer and Swanson (*1973*):

> . . . the SSD should never be employed for pairwise multiple comparisons . . . the TSD and SNK are clearly inferior in ability to detect real differences. Although the SSD, TSD, and SNK provide excellent protection against Type I errors, it is the authors' feeling that, in evaluation of the various procedures, concern for ability to detect real differences should receive a high priority . . . the FSD1 procedure also appears to stress protection against Type I errors at the expense of sensitivity . . . it also seems reasonable not to recommend procedures which unduly deemphasize protection against Type I errors. From this point of view, then, the ordinary LSD and perhaps the FSD3 can be eliminated from consideration; in addition, their sensitivities to real differences are not appreciably greater than those of the FSD2, BSD, BET, and MRT. These latter four procedures thus constitute a group from which the consulting statistician or experimenter might generally make a choice . . . while the MRT often produces a lower frequency of Type I errors, the other three are generally more sensitive in detecting real differences . . . dependence of the critical value on the observed analysis of variance F value is more appealing than dependence on the number of treatments in the experiment. Since the BET is an improved and more exact version than the BSD, it seems reasonable to prefer the former . . . the procedure (BET) is easier to apply than the MRT . . . many subject matter researchers will find the FSD2 attractive because of its simplicity and the fact that they are already familiar with Student's t table.

Carmer and Swanson's final choice is thus between FSD2 and BET. Waller and Duncan (*1969*) claim that the similarity in performance between the FSD2 and BET says a lot for BET, but as Carmer and Swanson point out, it is just as reasonable to claim that this similarity speaks a lot for the FSD2.

Thomas (*1974*) compared "seven methods of pairwise comparisons and four for constructing simultaneous sets of confidence limits. The general conclusions are that Duncan's multiple range test is the best method of those considered for the former and the Bonferroni t-based limits for the latter."

We mentioned at the beginning of this chapter that one main difficulty in comparing the procedures is due to the different kinds of Type I error rates used. Comparing one procedure using a 5% comparisonwise Type I error rate with another procedure using a 5% experimentwise Type I error rate is almost like comparing oranges with bananas. As Einot and Gabriel (*1975*) pointed out, any observed difference in the performance of the two procedures is more likely to be due to the different Type I error probabilities than to the techniques used. Therefore, one should force all procedures to have the same experimentwise (or comparisonwise) Type I error rate and compare their powers, as in the Neyman-Pearson two-decision situations. With orthogonal contrasts and large numbers of degrees of freedom for error mean square, we have seen in Section 3.1 that for $t = 10$ treatments, say, a 5% experimentwise error rate corresponds to a .57% comparisonwise error rate, and a 5% comparisonwise error rate is equivalent to a 36.98% experimentwise error rate.

Einot and Gabriel (*1975*) studied the powers of multiple comparison procedures for fixed maximal experimentwise levels, and ". . . generally recommend the Tukey technique for its elegant simplicity and existent confidence bounds—its power is little below that of any other method. Simulation was for 3, 4, and 5 treatments: the conclusions might need modification for more treatments."

No doubt the reader will think that the last word has not been written on the choice of a multiple comparison procedure. (Some statisticians do not even believe in multiple comparisons. In his discussion of the review paper by O'Neill and Wetherill (*1971*), R. L. Plackett expressed his "view that much of the subject of multiple comparisons is essentially artificial," while J. A. Nelder went so far as stating that in his opinion "multiple comparison methods have no place at all in the interpretation of data.") In the final analysis, the choice will be subjective. To a very large extent, this choice will hinge on a choice between an experimentwise error rate (for which Tukey's HSD is the recommended procedure) and a comparisonwise error rate (for which Duncan's MRT is recommended). As mentioned earlier, the author's opinion is that in the majority of cases, the comparisonwise basis is more appropriate since one wrong inference usually does not make the other inferences in the same experiment meaningless. There is really not that much difference between the methods. We can remove or reduce objections to Duncan's MRT by requiring an initial significant overall F test or by taking Duncan's comparisonwise α to be 0.01 or 0.001. Similarily, we can remove or reduce objections to Tukey's HSD by taking Tukey's experimentwise α to be 0.10 or 0.25, but, as Einot and Gabriel wondered, it may be that "it does not seem scientifically respectable to work explicitly with a level of 0.25."

The choice of the kind of Type I error rates is bypassed altogether in the Waller-Duncan Bayesian k-ratio LSD rule. It also has the extremely appealing feature that the observed F value is used in the calculation of the LSD. With a large F (of 3.0 and above, indicating strong evidence of existence of differences), the test behaves like the comparisonwise procedures (Duncan's MRT and Fisher's LSD) with good power properties, while for a small F, it becomes conservative with good protection against Type I error, as in the Tukey HSD procedure. It is as if the choice between a comparisonwise and an experimentwise error rate is taken out of the experimenter's hands and is determined by the experiment itself (the experimental F value). "In this way the decision theoretic rule enjoys the advantages of both comparisonwise and experimentwise α rules without their disadvantages." (Dixon and Duncan *1975*, p. 822). This procedure will become more popular in the future, especially if more extensive tables become available.

*TABLE A.—Two-sided (100 α/m)% points of student's t-distribution with ν degrees of freedom**

ν \ m	2	3	4	5	6	7	8	9	10	15	20	25	30	35	40	45	50
								$\alpha = .05$									
5	3.17	3.54	3.81	4.04	4.22	4.38	4.53	4.66	4.78	5.25	5.60	5.89	6.15	6.36	6.56	6.70	6.86
7	2.84	3.13	3.34	3.50	3.64	3.76	3.86	3.95	4.03	4.36	4.59	4.78	4.95	5.09	5.21	5.31	5.40
10	2.64	2.87	3.04	3.17	3.28	3.37	3.45	3.52	3.58	3.83	4.01	4.15	4.27	4.37	4.45	4.53	4.59
12	2.56	2.78	2.94	3.06	3.15	3.24	3.31	3.37	3.43	3.65	3.80	3.93	4.04	4.13	4.20	4.26	4.32
15	2.49	2.69	2.84	2.95	3.04	3.11	3.18	3.24	3.29	3.48	3.62	3.74	3.82	3.90	3.97	4.02	4.07
20	2.42	2.61	2.75	2.85	2.93	3.00	3.06	3.11	3.16	3.33	3.46	3.55	3.63	3.70	3.76	3.80	3.85
24	2.39	2.58	2.70	2.80	2.88	2.94	3.00	3.05	3.09	3.26	3.38	3.47	3.54	3.61	3.66	3.70	3.74
30	2.36	2.54	2.66	2.75	2.83	2.89	2.94	2.99	3.03	3.19	3.30	3.39	3.46	3.52	3.57	3.61	3.65
40	2.33	2.50	2.62	2.71	2.78	2.84	2.89	2.93	2.97	3.12	3.23	3.31	3.38	3.43	3.48	3.51	3.55
60	2.30	2.47	2.58	2.66	2.73	2.79	2.84	2.88	2.92	3.06	3.16	3.24	3.30	3.34	3.39	3.42	3.46
120	2.27	2.43	2.54	2.62	2.68	2.74	2.79	2.83	2.86	2.99	3.09	3.16	3.22	3.27	3.31	3.34	3.37
∞	2.24	2.39	2.50	2.58	2.64	2.69	2.74	2.77	2.81	2.94	3.02	3.09	3.15	3.19	3.23	3.26	3.29
								$\alpha = .01$									
5	4.78	5.25	5.60	5.89	6.15	6.36	6.56	6.70	6.86	7.51	8.00	8.37	8.68	8.95	9.19	9.41	9.68
7	4.03	4.36	4.59	4.78	4.95	5.09	5.21	5.31	5.40	5.79	6.08	6.30	6.49	6.67	6.83	6.93	7.06
10	3.58	3.83	4.01	4.15	4.27	4.37	4.45	4.53	4.59	4.86	5.06	5.20	5.33	5.44	5.52	5.60	5.70
12	3.43	3.65	3.80	3.93	4.04	4.13	4.20	4.26	4.32	4.56	4.73	4.86	4.95	5.04	5.12	5.20	5.27
15	3.29	3.48	3.62	3.74	3.82	3.90	3.97	4.02	4.07	4.29	4.42	4.53	4.61	4.71	4.78	4.84	4.90
20	3.16	3.33	3.46	3.55	3.63	3.70	3.76	3.80	3.85	4.03	4.15	4.25	4.33	4.39	4.46	4.52	4.56
24	3.09	3.26	3.38	3.47	3.54	3.61	3.66	3.70	3.74	3.91	4.04	4.1†	4.2†	4.3†	4.3†	4.3†	4.4†
30	3.03	3.19	3.30	3.39	3.46	3.52	3.57	3.61	3.65	3.80	3.90	3.98	4.13	4.26	4.1†	4.2†	4.2†
40	2.97	3.12	3.23	3.31	3.38	3.43	3.48	3.51	3.55	3.70	3.79	3.88	3.93	3.97	4.01	4.1†	4.1†
60	2.92	3.06	3.16	3.24	3.30	3.34	3.39	3.42	3.46	3.59	3.69	3.76	3.81	3.84	3.89	3.93	3.97
120	2.86	2.99	3.09	3.16	3.22	3.27	3.31	3.34	3.37	3.50	3.58	3.64	3.69	3.73	3.77	3.80	3.83
∞	2.81	2.94	3.02	3.09	3.15	3.19	3.23	3.26	3.29	3.40	3.48	3.54	3.59	3.63	3.66	3.69	3.72

†Obtained by graphical interpolation.

Source: Reproduced from Olive Jean Dunn, Multiple Comparisons Among Means. Journal of the American Statistical Association, vol. 56 (1961), pp. 52–64, with the permission of the author and the editor.

*TABLE B.—Percentage points of the studentized range $q(\alpha;p,\nu)$**

$\alpha = .05$

ν \ p	2	3	4	5	6	7	8	9	10
1	17.97	26.98	32.82	37.08	40.41	43.12	45.40	47.36	49.07
2	6.085	8.331	9.798	10.88	11.74	12.44	13.03	13.54	13.99
3	4.501	5.910	6.825	7.502	8.037	8.478	8.853	9.177	9.462
4	3.927	5.040	5.757	6.287	6.707	7.053	7.347	7.602	7.826
5	3.635	4.602	5.218	5.673	6.033	6.330	6.582	6.802	6.995
6	3.461	4.339	4.896	5.305	5.628	5.895	6.122	6.319	6.493
7	3.344	4.165	4.681	5.060	5.359	5.606	5.815	5.998	6.158
8	3.261	4.041	4.529	4.886	5.167	5.399	5.597	5.767	5.918
9	3.199	3.949	4.415	4.756	5.024	5.244	5.432	5.595	5.739
10	3.151	3.877	4.327	4.654	4.912	5.124	5.305	5.461	5.599
11	3.113	3.820	4.256	4.574	4.823	5.028	5.202	5.353	5.487
12	3.082	3.773	4.199	4.508	4.751	4.950	5.119	5.265	5.395
13	3.055	3.735	4.151	4.453	4.690	4.885	5.049	5.192	5.318
14	3.033	3.702	4.111	4.407	4.639	4.829	4.990	5.131	5.254
15	3.014	3.674	4.076	4.367	4.595	4.782	4.940	5.077	5.198
16	2.998	3.649	4.046	4.333	4.557	4.741	4.897	5.031	5.150
17	2.984	3.628	4.020	4.303	4.524	4.705	4.858	4.991	5.108
18	2.971	3.609	3.997	4.277	4.495	4.673	4.824	4.956	5.071
19	2.960	3.593	3.977	4.253	4.469	4.465	4.794	4.924	5.038
20	2.950	3.578	3.958	4.232	4.445	4.620	4.768	4.896	5.008
24	2.919	3.532	3.901	4.166	4.373	4.541	4.684	4.807	4.915
30	2.888	3.486	3.845	4.102	4.302	4.464	4.602	4.720	4.824
40	2.858	3.442	3.791	4.039	4.232	4.389	4.521	4.635	4.735
60	2.829	3.399	3.737	3.977	4.163	4.314	4.441	4.550	4.646
120	2.800	3.356	3.685	3.917	4.096	4.241	4.363	4.468	4.560
∞	2.772	3.314	3.633	3.858	4.030	4.170	4.286	4.387	4.474

TABLE B.—Percentage points of the studentized range $q(\alpha;p,\nu)$—Continued*

$\alpha = .05$

ν \ p	11	12	13	14	15	16	17	18	19
1	50.59	51.96	53.20	54.33	55.36	56.32	57.22	58.04	58.83
2	14.39	14.75	15.08	15.38	15.65	15.91	16.14	16.37	16.57
3	9.717	9.946	10.15	10.35	10.53	10.69	10.84	10.98	11.11
4	8.027	8.208	8.373	8.525	8.664	8.794	8.914	9.028	9.134
5	7.168	7.324	7.466	7.596	7.717	7.828	7.932	8.030	8.122
6	6.649	6.789	6.917	7.034	7.143	7.244	7.338	7.426	7.508
7	6.302	6.431	6.550	6.658	6.759	6.852	6.939	7.020	7.097
8	6.054	6.175	6.287	6.389	6.483	6.571	6.653	6.729	6.802
9	5.867	5.983	6.089	6.186	6.276	6.359	6.437	6.510	6.579
10	5.722	5.833	5.935	6.028	6.114	6.194	6.269	6.339	6.405
11	5.605	5.713	5.811	5.901	5.984	6.062	6.134	6.202	6.265
12	5.511	5.615	5.710	5.798	5.878	5.953	6.023	6.089	6.151
13	5.431	5.533	5.625	5.711	5.789	5.862	5.931	5.995	6.055
14	5.364	5.463	5.554	5.637	5.714	5.786	5.852	5.915	5.974
15	5.306	5.404	5.493	5.574	5.649	5.720	5.785	5.846	5.904
16	5.256	5.352	5.439	5.520	5.593	5.662	5.727	5.786	5.843
17	5.212	5.307	5.392	5.471	5.544	5.612	5.675	5.734	5.790
18	5.174	5.267	5.352	5.429	5.501	5.568	5.630	5.688	5.743
19	5.140	5.231	5.315	5.391	5.462	5.528	5.589	5.647	5.701
20	5.108	5.199	5.282	5.357	5.427	5.493	5.553	5.610	5.663
24	5.012	5.099	5.179	5.251	5.319	5.381	5.439	5.494	5.545
30	4.917	5.001	5.077	5.147	5.211	5.271	5.327	5.379	5.429
40	4.824	4.904	4.977	5.044	5.106	5.163	5.216	5.266	5.313
60	4.732	4.808	4.878	4.942	5.001	5.056	5.107	5.154	5.199
120	4.641	4.714	4.781	4.842	4.898	4.950	4.998	5.044	5.086
∞	4.552	4.622	4.685	4.743	4.796	4.845	4.891	4.934	4.974

TABLE B.—Percentage points of the studentized range $q(\alpha;p,\nu)$—Continued*

$\alpha = .05$

ν \ p	20	22	24	26	28	30	32	34	36
1	59.56	60.91	62.12	63.22	64.23	65.15	66.01	66.81	67.56
2	16.77	17.13	17.45	17.75	18.02	18.27	18.50	18.72	18.92
3	11.24	11.47	11.68	11.87	12.05	12.21	12.36	12.50	12.63
4	9.233	9.418	9.584	9.736	9.875	10.00	10.12	10.23	10.34
5	8.208	8.368	8.512	8.643	8.764	8.875	8.979	9.075	9.165
6	7.587	7.730	7.861	7.979	8.088	8.189	8.283	8.370	8.452
7	7.170	7.303	7.423	7.533	7.634	7.728	7.814	7.895	7.972
8	6.870	6.995	7.109	7.212	7.307	7.395	7.477	7.554	7.625
9	6.644	6.763	6.871	6.970	7.061	7.145	7.222	7.295	7.363
10	6.467	6.582	6.686	6.781	6.868	6.948	7.023	7.093	7.159
11	6.326	6.436	6.536	6.628	6.712	6.790	6.863	6.930	6.994
12	6.209	6.317	6.414	6.503	6.585	6.660	6.731	6.796	6.858
13	6.112	6.217	6.312	6.398	6.478	6.551	6.620	6.684	6.744
14	6.029	6.132	6.224	6.309	6.387	6.459	6.526	6.588	6.647
15	5.958	6.059	6.149	6.233	6.309	6.379	6.445	6.506	6.564
16	5.897	5.995	6.084	6.166	6.241	6.310	6.374	6.434	6.491
17	5.842	5.940	6.027	6.107	6.181	6.249	6.313	6.372	6.427
18	5.794	5.890	5.977	6.055	6.128	6.195	6.258	6.316	6.371
19	5.752	5.846	5.932	6.009	6.081	6.147	6.209	6.267	6.321
20	5.714	5.807	5.891	5.968	6.039	6.104	6.165	6.222	6.275
24	5.594	5.683	5.764	5.838	5.906	5.968	6.027	6.081	6.132
30	5.475	5.561	5.638	5.709	5.774	5.833	5.889	5.941	5.990
40	5.358	5.439	5.513	5.581	5.642	5.700	5.753	5.803	5.849
60	5.241	5.319	5.389	5.453	5.512	5.566	5.617	5.664	5.708
120	5.126	5.200	5.266	5.327	5.382	5.434	5.481	5.526	5.568
∞	5.012	5.081	5.144	5.201	5.253	5.301	5.346	5.388	5.427

TABLE B.—Percentage points of the studentized range $q(\alpha;p,\nu)$—Continued*

$\alpha = .05$

ν \ p	38	40	50	60	70	80	90	100
1	68.26	68.92	71.73	73.97	75.82	77.40	78.77	79.98
2	19.11	19.28	20.05	20.66	21.16	21.59	21.96	22.29
3	12.75	12.87	13.36	13.76	14.08	14.36	14.61	14.82
4	10.44	10.53	10.93	11.24	11.51	11.73	11.92	12.09
5	9.250	9.330	9.674	9.949	10.18	10.38	10.54	10.69
6	8.529	8.601	8.913	9.163	9.370	9.548	9.702	9.839
7	8.043	8.110	8.400	8.632	8.824	8.989	9.133	9.261
8	7.693	7.756	8.029	8.248	8.430	8.586	8.722	8.843
9	7.428	7.488	7.749	7.958	8.132	8.281	8.410	8.526
10	7.220	7.279	7.529	7.730	7.897	8.041	8.166	8.276
11	7.053	7.110	7.352	7.546	7.708	7.847	7.968	8.075
12	6.916	6.970	7.205	7.394	7.552	7.687	7.804	7.909
13	6.800	6.854	7.083	7.267	7.421	7.552	7.667	7.769
14	6.702	6.754	6.979	7.159	7.309	7.438	7.550	7.650
15	6.618	6.669	6.888	7.065	7.212	7.339	7.449	7.546
16	6.544	6.594	6.810	6.984	7.128	7.252	7.360	7.457
17	6.479	6.529	6.741	6.912	7.054	7.176	7.283	7.377
18	6.422	6.471	6.680	6.848	6.989	7.109	7.213	7.307
19	6.371	6.419	6.626	6.792	6.930	7.048	7.152	7.244
20	6.325	6.373	6.576	6.740	6.877	6.994	7.097	7.187
24	6.181	6.226	6.421	6.579	6.710	6.822	6.920	7.008
30	6.037	6.080	6.267	6.417	6.543	6.650	6.744	6.827
40	5.893	5.934	6.112	6.255	6.375	6.477	6.566	6.645
60	5.750	5.789	5.958	6.093	6.206	6.303	6.387	6.462
120	5.607	5.644	5.802	5.929	6.035	6.126	6.205	6.275
∞	5.463	5.498	5.646	5.764	5.863	5.947	6.020	6.085

TABLE B.—Percentage points of the studentized range $q(\alpha;p,\nu)$—Continued*

$\alpha = .01$

ν \ p	2	3	4	5	6	7	8	9	10
1	90.03	135.0	164.3	185.6	202.2	215.8	227.2	237.0	245.6
2	14.04	19.02	22.29	24.72	26.63	28.20	29.53	30.68	31.69
3	8.261	10.62	12.17	13.33	14.24	15.00	15.64	16.20	16.69
4	6.512	8.120	9.173	9.958	10.58	11.10	11.55	11.93	12.27
5	5.702	6.976	7.804	8.421	8.913	9.321	9.669	9.972	10.24
6	5.243	6.331	7.033	7.556	7.973	8.318	8.613	8.869	9.097
7	4.949	5.919	6.543	7.005	7.373	7.679	7.939	8.166	8.368
8	4.746	5.635	6.204	6.625	6.960	7.237	7.474	7.681	7.863
9	4.596	5.428	5.957	6.348	6.658	6.915	7.134	7.325	7.495
10	4.482	5.270	5.769	6.136	6.428	6.669	6.875	7.055	7.213
11	4.392	5.146	5.621	5.970	6.247	6.476	6.672	6.842	6.992
12	4.320	5.046	5.502	5.836	6.101	6.321	6.507	6.670	6.814
13	4.260	4.964	5.404	5.727	5.981	6.192	6.372	6.528	6.667
14	4.210	4.895	5.322	5.634	5.881	6.085	6.258	6.409	6.543
15	4.168	4.836	5.252	5.556	5.796	5.994	6.162	6.309	6.439
16	4.131	4.786	5.192	5.489	5.722	5.915	6.079	6.222	6.349
17	4.099	4.742	5.140	5.430	5.659	5.847	6.007	6.147	6.270
18	4.071	4.703	5.094	5.379	5.603	5.788	5.944	6.081	6.201
19	4.046	4.670	5.054	5.334	5.554	5.735	5.889	6.022	6.141
20	4.024	4.639	5.018	5.294	5.510	5.688	5.839	5.970	6.087
24	3.956	4.546	4.907	5.168	5.374	5.542	5.685	5.809	5.919
30	3.889	4.455	4.799	5.048	5.242	5.401	5.536	5.653	5.756
40	3.825	4.367	4.696	4.931	5.114	5.265	5.392	5.502	5.599
60	3.762	4.282	4.595	4.818	4.991	5.133	5.253	5.356	5.447
120	3.702	4.200	4.497	4.709	4.872	5.005	5.118	5.214	5.299
∞	3.643	4.120	4.403	4.603	4.757	4.882	4.987	5.078	5.157

TABLE B.—Percentage points of the studentized range $q(\alpha;p,\nu)$—Continued*

$\alpha = .01$

ν \ p	11	12	13	14	15	16	17	18	19
1	253.2	260.0	266.2	271.8	277.0	281.8	286.3	290.4	294.3
2	32.59	33.40	34.13	34.81	35.43	36.00	36.53	37.03	37.50
3	17.13	17.53	17.89	18.22	18.52	18.81	19.07	19.32	19.55
4	12.57	12.84	13.09	13.32	13.53	13.73	13.91	14.08	14.24
5	10.48	10.70	10.89	11.08	11.24	11.40	11.55	11.68	11.81
6	9.301	9.485	9.653	9.808	9.951	10.08	10.21	10.32	10.43
7	8.548	8.711	8.860	8.997	9.124	9.242	9.353	9.456	9.554
8	8.027	8.176	8.312	8.436	8.552	8.659	8.760	8.854	8.943
9	7.647	7.784	7.910	8.025	8.132	8.232	8.325	8.412	8.495
10	7.356	7.485	7.603	7.712	7.812	7.906	7.993	8.076	8.153
11	7.128	7.250	7.362	7.465	7.560	7.649	7.732	7.809	7.883
12	6.943	7.060	7.167	7.265	7.356	7.441	7.520	7.594	7.665
13	6.791	6.903	7.006	7.101	7.188	7.269	7.345	7.417	7.485
14	6.664	6.772	6.871	6.962	7.047	7.126	7.199	7.268	7.333
15	6.555	6.660	6.757	6.845	6.927	7.003	7.074	7.142	7.204
16	6.462	6.564	6.658	6.744	6.823	6.898	6.967	7.032	7.093
17	6.381	6.480	6.572	6.656	6.734	6.806	6.873	6.937	6.997
18	6.310	6.407	6.497	6.579	6.655	6.725	6.792	6.854	6.912
19	6.247	6.342	6.430	6.510	6.585	6.654	6.719	6.780	6.837
20	6.191	6.285	6.371	6.450	6.523	6.591	6.654	6.714	6.771
24	6.017	6.106	6.186	6.261	6.330	6.394	6.453	6.510	6.563
30	5.849	5.932	6.008	6.078	6.143	6.203	6.259	6.311	6.361
40	5.686	5.764	5.835	5.900	5.961	6.017	6.069	6.119	6.165
60	5.528	5.601	5.667	5.728	5.785	5.837	5.886	5.931	5.974
120	5.375	5.443	5.505	5.562	5.614	5.662	5.708	5.750	5.790
∞	5.227	5.290	5.348	5.400	5.448	5.493	5.535	5.574	5.611

TABLE B.—Percentage points of the studentized range $q(\alpha;p,\nu)$—Continued*

$\alpha = .01$

ν \ p	20	22	24	26	28	30	32	34	36
1	298.0	304.7	310.8	316.3	321.3	326.0	330.3	334.3	338.0
2	37.95	38.76	39.49	40.15	40.76	41.32	41.84	42.33	42.78
3	19.77	20.17	20.53	20.86	21.16	21.44	21.70	21.95	22.17
4	14.40	14.68	14.93	15.16	15.37	15.57	15.75	15.92	16.08
5	11.93	12.16	12.36	12.54	12.71	12.87	13.02	13.15	13.28
6	10.54	10.73	10.91	11.06	11.21	11.34	11.47	11.58	11.69
7	9.646	9.815	9.970	10.11	10.24	10.36	10.47	10.58	10.67
8	9.027	9.182	9.322	9.450	9.569	9.678	9.779	9.874	9.964
9	8.573	8.717	8.847	8.966	9.075	9.177	9.271	9.360	9.443
10	8.226	8.361	8.483	8.595	8.698	8.794	8.883	8.966	9.044
11	7.952	8.080	8.196	8.303	8.400	8.491	8.575	8.654	8.728
12	7.731	7.853	7.964	8.066	8.159	8.246	8.327	8.402	8.473
13	7.548	7.665	7.772	7.870	7.960	8.043	8.121	8.193	8.262
14	7.395	7.508	7.611	7.705	7.792	7.873	7.948	8.018	8.084
15	7.264	7.374	7.474	7.566	7.650	7.728	7.800	7.869	7.932
16	7.152	7.258	7.356	7.445	7.527	7.602	7.673	7.739	7.802
17	7.053	7.158	7.253	7.340	7.420	7.493	7.563	7.627	7.687
18	6.968	7.070	7.163	7.247	7.325	7.398	7.465	7.528	7.587
19	6.891	6.992	7.082	7.166	7.242	7.313	7.379	7.440	7.498
20	6.823	6.922	7.011	7.092	7.168	7.237	7.302	7.362	7.419
24	6.612	6.705	6.789	6.865	6.936	7.001	7.062	7.119	7.173
30	6.407	6.494	6.572	6.644	6.710	6.772	6.828	6.881	6.932
40	6.209	6.289	6.362	6.429	6.490	6.547	6.600	6.650	6.697
60	6.015	6.090	6.158	6.220	6.277	6.330	6.378	6.424	6.467
120	5.827	5.897	5.959	6.016	6.069	6.117	6.162	6.204	6.244
∞	5.645	5.709	5.766	5.818	5.866	5.911	5.952	5.990	6.026

TABLE B.—Percentage points of the studentized range $q(\alpha;p,\nu)$—Continued*

$\alpha = .01$

ν \ p	38	40	50	60	70	80	90	100
1	341.5	344.8	358.9	370.1	379.4	387.3	394.1	400.1
2	43.21	43.61	45.33	46.70	47.83	48.80	49.64	50.38
3	22.39	22.59	23.45	24.13	24.71	25.19	25.62	25.99
4	16.23	16.37	16.98	17.46	17.86	18.20	18.50	18.77
5	13.40	13.52	14.00	14.39	14.72	14.99	15.23	15.45
6	11.80	11.90	12.31	12.65	12.92	13.16	13.37	13.55
7	10.77	10.85	11.23	11.52	11.77	11.99	12.17	12.34
8	10.05	10.13	10.47	10.75	10.97	11.17	11.34	11.49
9	9.521	9.594	9.912	10.17	10.38	10.57	10.73	10.87
10	9.117	9.187	9.486	9.726	9.927	10.10	10.25	10.39
11	8.798	8.864	9.148	9.377	9.568	9.732	9.875	10.00
12	8.539	8.603	8.875	9.094	9.277	9.434	9.571	9.693
13	8.326	8.387	8.648	8.859	9.035	9.187	9.318	9.436
14	8.146	8.204	8.457	8.661	8.832	8.978	9.106	9.219
15	7.992	8.049	8.295	8.492	8.658	8.800	8.924	9.035
16	7.860	7.916	8.154	8.347	8.507	8.646	8.767	8.874
17	7.745	7.799	8.031	8.219	8.377	8.511	8.630	8.735
18	7.643	7.696	7.924	8.107	8.261	8.393	8.508	8.611
19	7.553	7.605	7.828	8.008	8.159	8.288	8.401	8.502
20	7.473	7.523	7.742	7.919	8.067	8.194	8.305	8.404
24	7.223	7.270	7.476	7.642	7.780	7.900	8.004	8.097
30	6.978	7.023	7.215	7.370	7.500	7.611	7.709	7.796
40	6.740	6.782	6.960	7.104	7.225	7.328	7.419	7.500
60	6.507	6.546	6.710	6.843	6.954	7.050	7.133	7.207
120	6.281	6.316	6.467	6.588	6.689	6.776	6.852	6.919
∞	6.060	6.092	6.228	6.338	6.429	6.507	6.575	6.636

Source: Reproduced from H. Leon Harter, Order Statistics and Their Use in Testing and Estimation, vol. 1 (1970), pp. 623–661, U.S. Government Printing Office, Washington, D.C., with the permission of the author.

TABLE C.—*Critical values for Duncan's Multiple Range Test*

$\alpha = .05$

ν \ p	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97
2	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085
3	4.501	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516
4	3.927	4.013	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033
5	3.635	3.749	3.797	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814
6	3.461	3.587	3.649	3.680	3.694	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697
7	3.344	3.477	3.548	3.588	3.611	3.622	3.626	3.626	3.626	3.626	3.626	3.626	3.626	3.626	3.626	3.626	3.626	3.626
8	3.261	3.399	3.475	3.521	3.549	3.566	3.575	3.579	3.579	3.579	3.579	3.579	3.579	3.579	3.579	3.579	3.579	3.579
9	3.199	3.339	3.420	3.470	3.502	3.523	3.536	3.544	3.547	3.547	3.547	3.547	3.547	3.547	3.547	3.547	3.547	3.547
10	3.151	3.293	3.376	3.430	3.465	3.489	3.505	3.516	3.522	3.525	3.526	3.526	3.526	3.526	3.526	3.526	3.526	3.526
11	3.113	3.256	3.342	3.397	3.435	3.462	3.480	3.493	3.501	3.506	3.509	3.510	3.510	3.510	3.510	3.510	3.510	3.510
12	3.082	3.225	3.313	3.370	3.410	3.439	3.459	3.474	3.484	3.491	3.496	3.498	3.499	3.499	3.499	3.499	3.499	3.499
13	3.055	3.200	3.289	3.348	3.389	3.419	3.442	3.458	3.470	3.478	3.484	3.488	3.490	3.490	3.490	3.490	3.490	3.490
14	3.033	3.178	3.268	3.329	3.372	3.403	3.426	3.444	3.457	3.467	3.474	3.479	3.482	3.484	3.484	3.485	3.485	3.485
15	3.014	3.160	3.250	3.312	3.356	3.389	3.413	3.432	3.446	3.457	3.465	3.471	3.476	3.478	3.480	3.481	3.481	3.481
16	2.998	3.144	3.235	3.298	3.343	3.376	3.402	3.422	3.437	3.449	3.458	3.465	3.470	3.473	3.477	3.478	3.478	3.478
17	2.984	3.130	3.222	3.285	3.331	3.366	3.392	3.412	3.429	3.441	3.451	3.459	3.465	3.469	3.473	3.475	3.476	3.476
18	2.971	3.118	3.210	3.274	3.321	3.356	3.383	3.405	3.421	3.435	3.445	3.454	3.460	3.465	3.470	3.472	3.474	3.474
19	2.960	3.107	3.199	3.264	3.311	3.347	3.375	3.397	3.415	3.429	3.440	3.449	3.456	3.462	3.467	3.470	3.472	3.473
20	2.950	3.097	3.190	3.255	3.303	3.339	3.368	3.391	3.409	3.424	3.436	3.445	3.453	3.459	3.464	3.467	3.470	3.472
24	2.919	3.066	3.160	3.226	3.276	3.315	3.345	3.370	3.390	3.406	3.420	3.432	3.441	3.449	3.456	3.461	3.465	3.469
30	2.888	3.035	3.131	3.199	3.250	3.290	3.322	3.349	3.371	3.389	3.405	3.418	3.430	3.439	3.447	3.454	3.460	3.466
40	2.858	3.006	3.102	3.171	3.224	3.266	3.300	3.328	3.352	3.373	3.390	3.405	3.418	3.429	3.439	3.448	3.456	3.463
60	2.829	2.976	3.073	3.143	3.198	3.241	3.277	3.307	3.333	3.355	3.374	3.391	3.406	3.419	3.431	3.442	3.451	3.460
120	2.800	2.947	3.045	3.116	3.172	3.217	3.254	3.287	3.314	3.337	3.359	3.377	3.394	3.409	3.423	3.435	3.446	3.457
∞	2.772	2.918	3.017	3.089	3.146	3.193	3.232	3.265	3.294	3.320	3.343	3.363	3.382	3.399	3.414	3.428	3.442	3.454

TABLE C.—Critical values for Duncan's Multiple Range Test—Continued

$\alpha = .05$

ν \ p	20	22	24	26	28	30	32	34	36	38	40	50	60	70	80	90	100
1	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97	17.97
2	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085	6.085
3	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516	4.516
4	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033	4.033
5	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814	3.814
6	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697	3.697
7	3.626	3.626	3.626	3.626	3.626	3.626	3.626	3.626	3.626	3.626	3.626	3.626	3.626	3.626	3.626	3.626	3.626
8	3.579	3.579	3.579	3.579	3.579	3.579	3.579	3.579	3.579	3.579	3.579	3.579	3.579	3.579	3.579	3.579	3.579
9	3.547	3.547	3.547	3.547	3.547	3.547	3.547	3.547	3.547	3.547	3.547	3.547	3.547	3.547	3.547	3.547	3.547
10	3.526	3.526	3.526	3.526	3.526	3.526	3.526	3.526	3.526	3.526	3.526	3.526	3.526	3.526	3.526	3.526	3.526
11	3.510	3.510	3.510	3.510	3.510	3.510	3.510	3.510	3.510	3.510	3.510	3.510	3.510	3.510	3.510	3.510	3.510
12	3.499	3.499	3.499	3.499	3.499	3.499	3.499	3.499	3.499	3.499	3.499	3.499	3.499	3.499	3.499	3.499	3.499
13	3.490	3.490	3.490	3.490	3.490	3.490	3.490	3.490	3.490	3.490	3.490	3.490	3.490	3.490	3.490	3.490	3.490
14	3.485	3.485	3.485	3.485	3.485	3.485	3.485	3.485	3.485	3.485	3.485	3.485	3.485	3.485	3.485	3.485	3.485
15	3.481	3.481	3.481	3.481	3.481	3.481	3.481	3.481	3.481	3.481	3.481	3.481	3.481	3.481	3.481	3.481	3.481
16	3.478	3.478	3.478	3.478	3.478	3.478	3.478	3.478	3.478	3.478	3.478	3.478	3.478	3.478	3.478	3.478	3.478
17	3.476	3.476	3.476	3.476	3.476	3.476	3.476	3.476	3.476	3.476	3.476	3.476	3.476	3.476	3.476	3.476	3.476
18	3.474	3.474	3.474	3.474	3.474	3.474	3.474	3.474	3.474	3.474	3.474	3.474	3.474	3.474	3.474	3.474	3.474
19	3.474	3.474	3.474	3.474	3.474	3.474	3.474	3.474	3.474	3.474	3.474	3.474	3.474	3.474	3.474	3.474	3.474
20	3.473	3.474	3.474	3.474	3.474	3.474	3.474	3.474	3.474	3.474	3.474	3.474	3.474	3.474	3.474	3.474	3.474
24	3.471	3.475	3.477	3.477	3.477	3.477	3.477	3.477	3.477	3.477	3.477	3.477	3.477	3.477	3.477	3.477	3.477
30	3.470	3.477	3.481	3.484	3.486	3.486	3.486	3.486	3.486	3.486	3.486	3.486	3.486	3.486	3.486	3.486	3.486
40	3.469	3.479	3.486	3.492	3.497	3.500	3.503	3.504	3.504	3.504	3.504	3.504	3.504	3.504	3.504	3.504	3.504
60	3.467	3.481	3.492	3.501	3.509	3.515	3.521	3.525	3.529	3.531	3.534	3.537	3.537	3.537	3.537	3.537	3.537
120	3.466	3.483	3.498	3.511	3.522	3.532	3.541	3.548	3.555	3.561	3.566	3.585	3.596	3.600	3.601	3.601	3.601
∞	3.466	3.486	3.505	3.522	3.536	3.550	3.562	3.574	3.584	3.594	3.603	3.640	3.668	3.690	3.708	3.722	3.735

TABLE C.—Critical values for Duncan's Multiple Range Test—Continued

$\alpha = .01$

ν \ p	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03
2	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04
3	8.261	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321
4	6.512	6.677	6.740	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756
5	5.702	5.893	5.989	6.040	6.065	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074
6	5.243	5.439	5.549	5.614	5.655	5.680	5.694	5.701	5.703	5.703	5.703	5.703	5.703	5.703	5.703	5.703	5.703	5.703
7	4.949	5.145	5.260	5.334	5.383	5.416	5.439	5.454	5.464	5.470	5.472	5.472	5.472	5.472	5.472	5.472	5.472	5.472
8	4.746	4.939	5.057	5.135	5.189	5.227	5.256	5.276	5.291	5.302	5.309	5.314	5.316	5.317	5.317	5.317	5.317	5.317
9	4.596	4.787	4.906	4.986	5.043	5.086	5.118	5.142	5.160	5.174	5.185	5.193	5.199	5.203	5.205	5.206	5.206	5.206
10	4.482	4.671	4.790	4.871	4.931	4.975	5.010	5.037	5.058	5.074	5.088	5.098	5.106	5.112	5.117	5.120	5.122	5.124
11	4.392	4.579	4.697	4.780	4.841	4.887	4.924	4.952	4.975	4.994	5.009	5.021	5.031	5.039	5.045	5.050	5.054	5.057
12	4.320	4.504	4.622	4.706	4.767	4.815	4.852	4.883	4.907	4.927	4.944	4.958	4.969	4.978	4.986	4.993	4.998	5.002
13	4.260	4.442	4.560	4.644	4.706	4.755	4.793	4.824	4.850	4.872	4.889	4.904	4.917	4.928	4.937	4.944	4.950	4.956
14	4.210	4.391	4.508	4.591	4.654	4.704	4.743	4.775	4.802	4.824	4.843	4.859	4.872	4.884	4.894	4.902	4.910	4.916
15	4.168	4.347	4.463	4.547	4.610	4.660	4.700	4.733	4.760	4.783	4.803	4.820	4.834	4.846	4.857	4.866	4.874	4.881
16	4.131	4.309	4.425	4.509	4.572	4.622	4.663	4.696	4.724	4.748	4.768	4.786	4.800	4.813	4.825	4.835	4.844	4.851
17	4.099	4.275	4.391	4.475	4.539	4.589	4.630	4.664	4.693	4.717	4.738	4.756	4.771	4.785	4.797	4.807	4.816	4.824
18	4.071	4.246	4.362	4.445	4.509	4.560	4.601	4.635	4.664	4.689	4.711	4.729	4.745	4.759	4.772	4.783	4.792	4.801
19	4.046	4.220	4.335	4.419	4.483	4.534	4.575	4.610	4.639	4.665	4.686	4.705	4.722	4.736	4.749	4.761	4.771	4.780
20	4.024	4.197	4.312	4.395	4.459	4.510	4.552	4.587	4.617	4.642	4.664	4.684	4.701	4.716	4.729	4.741	4.751	4.761
24	3.956	4.126	4.239	4.322	4.386	4.437	4.480	4.516	4.546	4.573	4.596	4.616	4.634	4.651	4.665	4.678	4.690	4.700
30	3.889	4.056	4.168	4.250	4.314	4.366	4.409	4.445	4.477	4.504	4.528	4.550	4.569	4.586	4.601	4.615	4.628	4.640
40	3.825	3.988	4.098	4.180	4.244	4.296	4.339	4.376	4.408	4.436	4.461	4.483	4.503	4.521	4.537	4.553	4.566	4.579
60	3.762	3.922	4.031	4.111	4.174	4.226	4.270	4.307	4.340	4.368	4.394	4.417	4.438	4.456	4.474	4.490	4.504	4.518
120	3.702	3.858	3.965	4.044	4.107	4.158	4.202	4.239	4.272	4.301	4.327	4.351	4.372	4.392	4.410	4.426	4.442	4.456
∞	3.643	3.796	3.900	3.978	4.040	4.091	4.135	4.172	4.205	4.235	4.261	4.285	4.307	4.327	4.345	4.363	4.379	4.394

TABLE C.—Critical values for Duncan's Multiple Range Test—Continued

$\alpha = .01$

ν \ p	20	22	24	26	28	30	32	34	36	38	40	50	60	70	80	90	100
1	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03	90.03
2	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04	14.04
3	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321	8.321
4	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756	6.756
5	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074	6.074
6	5.703	5.703	5.703	5.703	5.703	5.703	5.703	5.703	5.703	5.703	5.703	5.703	5.703	5.703	5.703	5.703	5.703
7	5.472	5.472	5.472	5.472	5.472	5.472	5.472	5.472	5.472	5.472	5.472	5.472	5.472	5.472	5.472	5.472	5.472
8	5.317	5.317	5.317	5.317	5.317	5.317	5.317	5.317	5.317	5.317	5.317	5.317	5.317	5.317	5.317	5.317	5.317
9	5.206	5.206	5.206	5.206	5.206	5.206	5.206	5.206	5.206	5.206	5.206	5.206	5.206	5.206	5.206	5.206	5.206
10	5.124	5.124	5.124	5.124	5.124	5.124	5.124	5.124	5.124	5.124	5.124	5.124	5.124	5.124	5.124	5.124	5.124
11	5.059	5.061	5.061	5.061	5.061	5.061	5.061	5.061	5.061	5.061	5.061	5.061	5.061	5.061	5.061	5.061	5.061
12	5.006	5.010	5.011	5.011	5.011	5.011	5.011	5.011	5.011	5.011	5.011	5.011	5.011	5.011	5.011	5.011	5.011
13	4.960	4.966	4.970	4.972	4.972	4.972	4.972	4.972	4.972	4.972	4.972	4.972	4.972	4.972	4.972	4.972	4.972
14	4.921	4.929	4.935	4.938	4.940	4.940	4.940	4.940	4.940	4.940	4.940	4.940	4.940	4.940	4.940	4.940	4.940
15	4.887	4.897	4.904	4.909	4.912	4.914	4.914	4.914	4.914	4.914	4.914	4.914	4.914	4.914	4.914	4.914	4.914
16	4.858	4.869	4.877	4.883	4.887	4.890	4.892	4.892	4.892	4.892	4.892	4.892	4.892	4.892	4.892	4.892	4.892
17	4.832	4.844	4.853	4.860	4.865	4.869	4.872	4.873	4.874	4.874	4.874	4.874	4.874	4.874	4.874	4.874	4.874
18	4.808	4.821	4.832	4.839	4.846	4.850	4.854	4.856	4.857	4.858	4.858	4.858	4.858	4.858	4.858	4.858	4.858
19	4.788	4.802	4.812	4.821	4.828	4.833	4.838	4.841	4.843	4.844	4.845	4.845	4.845	4.845	4.845	4.845	4.845
20	4.769	4.786	4.795	4.805	4.813	4.818	4.823	4.827	4.830	4.832	4.833	4.833	4.833	4.833	4.833	4.833	4.833
24	4.710	4.727	4.741	4.752	4.762	4.770	4.777	4.783	4.788	4.791	4.794	4.802	4.802	4.802	4.802	4.802	4.802
30	4.650	4.669	4.685	4.699	4.711	4.721	4.730	4.738	4.744	4.750	4.755	4.772	4.777	4.777	4.777	4.777	4.777
40	4.591	4.611	4.630	4.645	4.659	4.671	4.682	4.692	4.700	4.708	4.715	4.740	4.754	4.761	4.764	4.764	4.764
60	4.530	4.553	4.573	4.591	4.607	4.620	4.633	4.645	4.655	4.665	4.673	4.707	4.730	4.745	4.755	4.761	4.765
120	4.469	4.494	4.516	4.535	4.552	4.568	4.583	4.596	4.609	4.619	4.630	4.673	4.703	4.727	4.745	4.759	4.770
∞	4.408	4.434	4.457	4.478	4.497	4.514	4.530	4.545	4.559	4.572	4.584	4.635	4.675	4.707	4.734	4.756	4.776

Source: Reproduced from H. Leon Harter, Critical Values for Duncan's New Multiple Range Test, Biometrics, vol. 16(1960), with the permission of the author and the editor.

TABLE D1.—Critical values of k-ratio t test (k = 100)

ν(denominator d.f. for F)

q(num. d.f. for F)	6	8	10	12	14	16	18	20	24	30	40	60	120
					F = 1.2 (a = .913, b = 2.449)								
2–6	*	*	*	*	*	*	*	*	*	*	*	*	*
8	2.91	2.94	2.96	2.97	2.98	2.99	2.99	2.99	3.00	3.00	3.00	3.00	3.00
10	2.93	2.98	3.01	3.04	3.05	3.06	3.07	3.08	3.09	3.10	3.10	3.11	3.12
12	2.95	3.01	3.05	3.08	3.10	3.12	3.13	3.14	3.16	3.17	3.19	3.20	3.21
14	2.96	3.03	3.08	3.12	3.14	3.16	3.18	3.19	3.21	3.23	3.25	3.27	3.29
16	2.97	3.05	3.11	3.15	3.18	3.20	3.22	3.24	3.26	3.28	3.31	3.33	3.36
20	2.99	3.08	3.14	3.19	3.23	3.26	3.28	3.30	3.33	3.37	3.40	3.44	3.47
40	3.02	3.13	3.22	3.29	3.35	3.39	3.43	3.47	3.52	3.58	3.64	3.72	3.79
100	3.04	3.17	3.28	3.36	3.44	3.50	3.55	3.59	3.67	3.76	3.86	3.98	4.11
∞	3.05	3.20	3.32	3.42	3.50	3.58	3.64	3.70	3.80	3.91	4.06	4.24	4.45
					F = 1.4 (a = .845, b = 1.871)								
2–4	*	*	*	*	*	*	*	*	*	*	*	*	*
6	2.85	2.84	2.83	2.82	2.82	2.81	2.80	2.80	2.79	2.78	2.77	2.75	2.74
8	2.88	2.89	2.90	2.90	2.90	2.89	2.89	2.89	2.88	2.88	2.87	2.86	2.85
10	2.90	2.93	2.94	2.95	2.95	2.96	2.96	2.96	2.95	2.95	2.95	2.94	2.93
12	2.92	2.95	2.98	2.99	3.00	3.00	3.01	3.01	3.01	3.01	3.01	3.00	2.99
14	2.93	2.97	3.00	3.02	3.03	3.04	3.04	3.05	3.05	3.06	3.06	3.05	3.05
16	2.94	2.99	3.02	3.04	3.06	3.07	3.08	3.08	3.09	3.09	3.10	3.10	3.09
20	2.95	3.01	3.05	3.08	3.10	3.11	3.12	3.13	3.14	3.15	3.16	3.16	3.16
40	2.98	3.06	3.12	3.16	3.19	3.22	3.24	3.25	3.28	3.30	3.31	3.32	3.32
100	2.99	3.09	3.16	3.22	3.26	3.29	3.32	3.34	3.38	3.41	3.43	3.45	3.42
∞	3.01	3.12	3.20	3.26	3.31	3.35	3.39	3.42	3.46	3.50	3.53	3.54	3.46
					F = 1.7 (a = .767, b = 1.558)								
2	*	*	*	*	*	*	*	*	*	*	*	*	*
4	*	*	*	*	*	2.61	2.59	2.58	2.56	2.54	2.52	2.50	2.48
6	2.82	2.79	2.76	2.74	2.72	2.71	2.70	2.69	2.67	2.65	2.63	2.61	2.58
8	2.84	2.83	2.81	2.80	2.78	2.77	2.76	2.75	2.74	2.72	2.70	2.68	2.65
10	2.86	2.86	2.85	2.84	2.83	2.82	2.81	2.80	2.79	2.77	2.75	2.73	2.70
12	2.87	2.88	2.88	2.87	2.86	2.85	2.84	2.84	2.82	2.81	2.79	2.76	2.73
14	2.88	2.90	2.90	2.89	2.89	2.88	2.87	2.86	2.85	2.83	2.81	2.79	2.75
16	2.89	2.91	2.91	2.91	2.90	2.90	2.89	2.89	2.87	2.86	2.84	2.81	2.77
20	2.90	2.93	2.93	2.94	2.93	2.93	2.92	2.92	2.91	2.89	2.87	2.84	2.80
40	2.93	2.97	2.99	3.00	3.00	3.00	3.00	2.99	2.98	2.97	2.94	2.89	2.83
100	2.94	2.99	3.02	3.04	3.05	3.05	3.05	3.05	3.04	3.02	2.98	2.92	2.83
∞	2.95	3.01	3.05	3.07	3.08	3.09	3.09	3.08	3.07	3.05	3.01	2.93	2.81
					F = 2.0 (a = .707, b = 1.414)								
2	*	*	*	*	*	*	*	*	*	*	*	*	*
4	2.74	2.67	2.63	2.59	2.56	2.54	2.52	2.51	2.49	2.46	2.44	2.41	2.39
6	2.79	2.74	2.70	2.67	2.64	2.62	2.60	2.59	2.57	2.54	2.52	2.49	2.46
8	2.81	2.77	2.74	2.71	2.69	2.67	2.65	2.64	2.62	2.59	2.56	2.53	2.49
10	2.83	2.80	2.77	2.74	2.72	2.70	2.69	2.67	2.65	2.62	2.59	2.56	2.52
12	2.84	2.82	2.79	2.77	2.75	2.73	2.71	2.70	2.67	2.64	2.61	2.57	2.53
14	2.85	2.83	2.81	2.79	2.77	2.75	2.73	2.72	2.69	2.66	2.63	2.59	2.54
16	2.85	2.84	2.82	2.80	2.78	2.76	2.74	2.73	2.70	2.67	2.64	2.59	2.54
20	2.86	2.85	2.84	2.82	2.80	2.78	2.77	2.75	2.72	2.69	2.65	2.61	2.55
40	2.88	2.89	2.88	2.86	2.85	2.83	2.81	2.80	2.77	2.73	2.68	2.62	2.55
100	2.89	2.91	2.90	2.89	2.88	2.86	2.84	2.82	2.79	2.75	2.69	2.62	2.53
∞	2.90	2.92	2.92	2.91	2.90	2.88	2.86	2.85	2.81	2.76	2.69	2.61	2.52

See footnotes at end of table.

TABLE D1.—Critical values of k-ratio t test (k = 100)—Continued

ν(denominator d.f. for F)

q(num. d.f. for F)	6	8	10	12	14	16	18	20	24	30	40	60	120
					F = 2.4 (a = .645, b = 1.309)								
2	*	*	*	*	*	*	*	*	*	*	*	*	2.18
4	2.71	2.63	2.57	2.53	2.49	2.47	2.44	2.43	2.40	2.37	2.34	2.31	2.28
6	2.75	2.68	2.63	2.58	2.55	2.52	2.50	2.48	2.46	2.42	2.39	2.36	2.32
8	2.77	2.71	2.66	2.62	2.59	2.56	2.54	2.52	2.49	2.45	2.42	2.38	2.34
10	2.79	2.73	2.68	2.64	2.61	2.58	2.56	2.54	2.50	2.47	2.43	2.39	2.34
12	2.79	2.74	2.70	2.66	2.62	2.60	2.57	2.55	2.52	2.48	2.44	2.39	2.35
14	2.80	2.75	2.71	2.67	2.64	2.61	2.58	2.56	2.53	2.49	2.44	2.40	2.35
16	2.81	2.76	2.72	2.68	2.65	2.62	2.59	2.57	2.53	2.49	2.45	2.40	2.34
20	2.82	2.77	2.73	2.69	2.66	2.63	2.60	2.58	2.54	2.50	2.45	2.40	2.34
40	2.83	2.80	2.76	2.72	2.69	2.66	2.63	2.60	2.56	2.51	2.46	2.39	2.33
100	2.84	2.81	2.78	2.74	2.71	2.67	2.64	2.62	2.57	2.51	2.45	2.39	2.32
∞	2.85	2.83	2.79	2.76	2.72	2.68	2.65	2.62	2.57	2.51	2.45	2.38	2.31
					F = 3.0 (a = .577, b = 1.225)								
2	*	*	2.41	2.36	2.32	2.29	2.27	2.25	2.22	2.20	2.17	2.14	2.11
4	2.68	2.57	2.50	2.45	2.41	2.38	2.35	2.33	2.30	2.27	2.24	2.20	2.17
6	2.71	2.61	2.54	2.49	2.44	2.41	2.39	2.36	2.33	2.29	2.26	2.22	2.18
8	2.72	2.63	2.56	2.51	2.47	2.43	2.40	2.38	2.34	2.31	2.27	2.22	2.18
10	2.74	2.65	2.58	2.52	2.48	2.44	2.41	2.39	2.35	2.31	2.27	2.22	2.18
12	2.74	2.66	2.59	2.53	2.49	2.45	2.42	2.40	2.36	2.31	2.27	2.22	2.18
14	2.75	2.66	2.60	2.54	2.49	2.46	2.43	2.40	2.36	2.32	2.27	2.22	2.17
16	2.75	2.67	2.60	2.55	2.50	2.46	2.43	2.40	2.36	2.32	2.27	2.22	2.17
20	2.76	2.68	2.61	2.55	2.51	2.47	2.43	2.41	2.36	2.32	2.27	2.22	2.17
40	2.77	2.70	2.63	2.57	2.52	2.48	2.44	2.41	2.37	2.32	2.26	2.21	2.16
100	2.78	2.71	2.64	2.58	2.53	2.49	2.45	2.42	2.37	2.31	2.26	2.21	2.16
∞	2.79	2.71	2.65	2.59	2.53	2.49	2.45	2.42	2.37	2.31	2.26	2.20	2.15
					F = 4.0 (a = .500, b = 1.155)								
2	2.58	2.44	2.35	2.29	2.25	2.22	2.20	2.18	2.15	2.12	2.09	2.06	2.03
4	2.63	2.50	2.41	2.35	2.30	2.27	2.24	2.22	2.18	2.15	2.12	2.08	2.05
6	2.65	2.52	2.43	2.37	2.32	2.28	2.25	2.23	2.19	2.16	2.12	2.08	2.04
10	2.67	2.55	2.46	2.39	2.34	2.30	2.26	2.24	2.20	2.16	2.12	2.08	2.04
20	2.69	2.57	2.47	2.40	2.35	2.30	2.27	2.24	2.20	2.15	2.11	2.07	2.03
∞	2.71	2.59	2.49	2.42	2.36	2.31	2.27	2.24	2.19	2.15	2.11	2.06	2.02
					F = 6.0 (a= .408, b = 1.095)								
2	2.53	2.37	2.27	2.21	2.16	2.13	2.10	2.08	2.05	2.02	1.99	1.96	1.93
4	2.56	2.40	2.30	2.23	2.18	2.14	2.12	2.09	2.06	2.02	1.99	1.96	1.93
6	2.58	2.42	2.31	2.24	2.19	2.15	2.12	2.09	2.06	2.02	1.99	1.95	1.92
10	2.59	2.43	2.32	2.24	2.19	2.15	2.12	2.09	2.06	2.02	1.99	1.95	1.92
20	2.60	2.44	2.32	2.25	2.19	2.15	2.12	2.09	2.05	2.02	1.98	1.95	1.92
∞	2.61	2.44	2.33	2.25	2.19	2.15	2.12	2.09	2.05	2.02	1.98	1.95	1.92

See footnotes at end of table.

TABLE D1.—Critical values of k-ratio t test (k = 100)—Continued

ν(denominator d.f. for F)

q(num. d.f. for F)	6	8	10	12	14	16	18	20	24	30	40	60	120
						F = 10.0 (a = .316, b = 1.054)							
2	2.48	2.30	2.19	2.12	2.07	2.04	2.01	1.99	1.96	1.93	1.90	1.87	1.85
4	2.49	2.31	2.20	2.13	2.08	2.04	2.01	1.99	1.96	1.93	1.90	1.87	1.84
6	2.50	2.31	2.20	2.13	2.08	2.04	2.01	1.99	1.96	1.93	1.90	1.87	1.84
10–∞	2.51	2.32	2.20	2.13	2.08	2.04	2.01	1.99	1.96	1.93	1.90	1.87	1.84
						F = 25.0(a = .200, b = 1.021)							
2-4	2.40	2.20	2.10	2.03	1.99	1.95	1.93	1.91	1.88	1.86	1.83	1.80	1.78
6-∞	2.41	2.21	2.10	2.03	1.99	1.95	1.93	1.91	1.88	1.86	1.83	1.80	1.78
						F = ∞ (a = 0, b = 1)							
2-∞	2.33	2.13	2.03	1.97	1.93	1.90	1.88	1.86	1.84	1.81	1.79	1.76	1.74

*All differences not significant. $a = 1/F^{½}$. $b = [F/(F-1)]^{½}$.

If $\nu=4$, t=2.83 for all q and F satisfying $F > 8.12/q$.

Source: Reproduced from Waller, Ray A., and Duncan, David B., A Bayes Rule for the Symmetric Multiple Comparisons Problem, Corrigenda, Journal of the American Statistical Association, vol. 67 (1972), with permission of author and publisher.

TABLE D2.—Critical values of k-ratio t test (k =500)

q (num. d.f. for F)	6	8	10	12	14	16	18	20	24	30	40	60	120
						ν (denominator d.f. for F)							
						F = 1.2 (a = .913, b = 2.449)							
2–16	*	*	*	*	*	*	*	*	*	*	*	*	*
20	4.70	4.82	4.89	*	*	*	*	*	*	*	*	*	*
40	4.75	4.91	5.03	5.12	5.20	5.25	5.30	5.34	5.41	5.48	5.55	5.61	5.67
100	4.79	4.98	5.13	5.25	5.34	5.43	5.50	5.56	5.65	5.76	5.89	6.02	6.13
∞	4.81	5.03	5.20	5.34	5.46	5.56	5.65	5.73	5.86	6.02	6.20	6.41	6.56
						F = 1.4 (a= .845, b = 1.871)							
2–14	*	*	*	*	*	*	*	*	*	*	*	*	*
16	4.61	4.66	4.68	4.69	4.69	4.69	4.69	4.68	4.67	4.65	4.62	4.58	4.53
20	4.64	4.70	4.73	4.75	4.76	4.77	4.77	4.76	4.76	4.74	4.72	4.68	4.62
40	4.68	4.78	4.85	4.89	4.92	4.94	4.96	4.96	4.97	4.97	4.95	4.90	4.81
∞	4.74	4.88	4.99	5.06	5.12	5.17	5.20	5.23	5.26	5.28	5.26	5.16	4.82
						F = 1.7 (a = .767, b = 1.558)							
2–8	*	*	*	*	*	*	*	*	*	*	*	*	*
10	*	*	*	*	*	*	*	*	*	4.08	4.02	3.95	3.87
12	4.50	4.46	4.42	4.38	4.34	4.30	4.27	4.24	4.19	4.14	4.07	3.99	3.90
20	4.55	4.54	4.52	4.49	4.46	4.43	4.40	4.37	4.32	4.26	4.18	4.08	3.95
40	4.59	4.61	4.61	4.60	4.57	4.55	4.52	4.49	4.44	4.36	4.26	4.12	3.93
∞	4.64	4.69	4.71	4.72	4.71	4.69	4.66	4.63	4.57	4.46	4.31	4.07	3.76
						F = 2.0 (a = .707, b = 1.414)							
2–6	*	*	*	*	*	*	*	*	*	*	*	*	*
8	*	*	*	*	*	3.98	3.93	3.89	3.83	3.76	3.69	3.60	3.51
10	4.41	4.31	4.22	4.15	4.08	4.03	3.98	3.94	3.88	3.80	3.72	3.63	3.53
20	4.48	4.41	4.34	4.27	4.21	4.16	4.10	4.06	3.98	3.89	3.78	3.65	3.51
40	4.51	4.47	4.41	4.35	4.29	4.23	4.17	4.12	4.03	3.92	3.78	3.62	3.44
∞	4.55	4.53	4.49	4.43	4.37	4.31	4.25	4.19	4.07	3.93	3.75	3.54	3.33
						F = 2.4 (a = .645, b = 1.309)							
2–4	*	*	*	*	*	*	*	*	*	*	*	*	*
6	*	*	*	*	3.77	3.71	3.65	3.61	3.54	3.47	3.39	3.30	3.22
8	4.31	4.14	4.01	3.91	3.83	3.76	3.70	3.66	3.58	3.50	3.41	3.32	3.22
10	4.33	4.18	4.05	3.95	3.87	3.79	3.73	3.68	3.60	3.51	3.42	3.31	3.21
20	4.39	4.26	4.14	4.04	3.95	3.87	3.80	3.74	3.64	3.53	3.41	3.28	3.15
∞	4.45	4.35	4.25	4.14	4.03	3.94	3.85	3.78	3.64	3.50	3.34	3.18	3.04
						F = 3.0 (a = .577, b = 1.225)							
2	*	*	*	*	*	*	*	*	*	*	*	*	*
4	*	*	*	*	*	3.43	3.38	3.33	3.26	3.19	3.12	3.04	2.97
6	4.19	3.95	3.79	3.66	3.56	3.49	3.43	3.37	3.30	3.21	3.13	3.04	2.95
10	4.24	4.02	3.85	3.72	3.62	3.53	3.46	3.40	3.31	3.21	3.12	3.02	2.92
20	4.28	4.08	3.91	3.77	3.65	3.56	3.48	3.41	3.31	3.20	3.09	2.98	2.87
∞	4.33	4.15	3.97	3.82	3.69	3.57	3.48	3.40	3.28	3.15	3.03	2.92	2.82

TABLE D2.—Critical values of k-ratio t test (k = 500)—Continued

ν (denominator d.f. for F)

q(num. d.f. for F)	6	8	10	12	14	16	18	20	24	30	40	60	120
						F = 4.0 (a = .500, b = 1.155)							
2	*	*	*	*	*	*	*	*	*	*	*	2.81	2.75
4	*	3.74	3.54	3.40	3.30	3.22	3.16	3.11	3.04	2.96	2.89	2.81	2.74
6	4.08	3.78	3.58	3.43	3.32	3.24	3.17	3.12	3.04	2.95	2.87	2.79	2.71
10	4.12	3.83	3.62	3.46	3.34	3.25	3.17	3.11	3.03	2.94	2.85	2.77	2.69
20	4.15	3.86	3.64	3.48	3.35	3.25	3.17	3.10	3.01	2.92	2.83	2.74	2.66
∞	4.19	3.90	3.67	3.49	3.35	3.24	3.15	3.09	2.99	2.89	2.80	2.72	2.65
						F = 6.0 (a = .408, b = 1.095)							
2	*	*	3.28	3.14	3.04	2.97	2.91	2.87	2.81	2.74	2.68	2.62	2.56
4	3.90	3.54	3.32	3.17	3.06	2.98	2.92	2.87	2.80	2.73	2.66	2.60	2.53
6	3.93	3.57	3.33	3.18	3.06	2.98	2.91	2.86	2.79	2.72	2.65	2.58	2.52
10	3.95	3.59	3.34	3.18	3.06	2.97	2.91	2.85	2.78	2.71	2.64	2.57	2.51
20	3.97	3.60	3.35	3.18	3.06	2.97	2.90	2.84	2.77	2.70	2.63	2.56	2.51
∞	3.99	3.62	3.36	3.18	3.05	2.96	2.89	2.83	2.76	2.69	2.62	2.56	2.50
						F = 10.0 (a = .316, b = 1.054)							
2	3.72	3.33	3.10	2.96	2.86	2.79	2.74	2.70	2.64	2.58	2.52	2.47	2.42
4	3.75	3.35	3.11	2.96	2.86	2.79	2.73	2.69	2.63	2.57	2.51	2.46	2.41
10	3.78	3.36	3.11	2.96	2.85	2.78	2.72	2.68	2.62	2.56	2.50	2.45	2.40
20	3.79	3.36	3.11	2.96	2.85	2.78	2.72	2.68	2.62	2.56	2.50	2.45	2.40
∞	3.80	3.37	3.11	2.95	2.85	2.77	2.72	2.67	2.61	2.56	2.50	2.45	2.40
						F = 25.0 (a = .200, b = 1.021)							
2	3.55	3.14	2.92	2.79	2.70	2.64	2.59	2.56	2.51	2.46	2.41	2.36	2.32
10	3.57	3.14	2.92	2.79	2.70	2.64	2.59	2.55	2.50	2.45	2.41	2.36	2.32
∞	3.57	3.14	2.92	2.78	2.70	2.63	2.59	2.55	2.50	2.45	2.41	2.36	2.32
						F = ∞ (a = 0, b = 1)							
2-∞	3.39	3.00	2.80	2.69	2.61	2.55	2.51	2.48	2.44	2.39	2.35	2.31	2.27

*All differences not significant. $a = 1/F^{\frac{1}{4}}$, $b = [F/(F - 1)]^{\frac{1}{4}}$.

If $\nu=4$, $t = 4.52$ for all q and F satisfying $F > 20.43/q$.

Source: Reproduced from Waller, Ray A., and Duncan, David B. A. Bayes Rule for the Symmetric Multiple Comparisons Problem, Corrigenda, Journal of the American Statistical Association, vol. 67 (1972), pp. 253-255, with the permission of the author and the publisher.

TABLE E.–100γ% points of the distribution of the largest absolute value of k uncorrelated Student t variates with ν degrees of freedom

ν	k	1	2	3	4	5	6	8	10	12	15	20
						γ=0.90						
3		2.353	2.989	3.369	3.637	3.844	4.011	4.272	4.471	4.631	4.823	5.066
4		2.132	2.662	2.976	3.197	3.368	3.506	3.722	3.887	4.020	4.180	4.383
5		2.015	2.491	2.769	2.965	3.116	3.239	3.430	3.576	3.694	3.837	4.018
6		1.943	2.385	2.642	2.822	2.961	3.074	3.249	3.384	3.493	3.624	3.790
7		1.895	2.314	2.556	2.725	2.856	2.962	3.127	3.253	3.355	3.478	3.635
8		1.860	2.262	2.494	2.656	2.780	2.881	3.038	3.158	3.255	3.373	3.522
9		1.833	2.224	2.447	2.603	2.723	2.819	2.970	3.086	3.179	3.292	3.436
10		1.813	2.193	2.410	2.562	2.678	2.771	2.918	3.029	3.120	3.229	3.368
11		1.796	2.169	2.381	2.529	2.642	2.733	2.875	2.984	3.072	3.178	3.313
12		1.782	2.149	2.357	2.501	2.612	2.701	2.840	2.946	3.032	3.136	3.268
15		1.753	2.107	2.305	2.443	2.548	2.633	2.765	2.865	2.947	3.045	3.170
20		1.725	2.065	2.255	2.386	2.486	2.567	2.691	2.786	2.863	2.956	3.073
25		1.708	2.041	2.226	2.353	2.450	2.528	2.648	2.740	2.814	2.903	3.016
30		1.697	2.025	2.207	2.331	2.426	2.502	2.620	2.709	2.781	2.868	2.978
40		1.684	2.006	2.183	2.305	2.397	2.470	2.585	2.671	2.741	2.825	2.931
60		1.671	1.986	2.160	2.278	2.368	2.439	2.550	2.634	2.701	2.782	2.884
						γ = 0.95						
3		3.183	3.960	4.430	4.764	5.023	5.233	5.562	5.812	6.015	6.259	6.567
4		2.777	3.382	3.745	4.003	4.203	4.366	4.621	4.817	4.975	5.166	5.409
5		2.571	3.091	3.399	3.619	3.789	3.928	4.145	4.312	4.447	4.611	4.819
6		2.447	2.916	3.193	3.389	3.541	3.664	3.858	4.008	4.129	4.275	4.462
7		2.365	2.800	3.056	3.236	3.376	3.489	3.668	3.805	3.916	4.051	4.223
8		2.306	2.718	2.958	3.128	3.258	3.365	3.532	3.660	3.764	3.891	4.052
9		2.262	2.657	2.885	3.046	3.171	3.272	3.430	3.552	3.651	3.770	3.923
10		2.228	2.609	2.829	2.984	3.103	3.199	3.351	3.468	3.562	3.677	3.823
11		2.201	2.571	2.784	2.933	3.048	3.142	3.288	3.400	3.491	3.602	3.743
12		2.179	2.540	2.747	2.892	3.004	3.095	3.236	3.345	3.433	3.541	3.677
15		2.132	2.474	2.669	2.805	2.910	2.994	3.126	3.227	3.309	3.409	3.536
20		2.086	2.411	2.594	2.722	2.819	2.898	3.020	3.114	3.190	3.282	3.399
25		2.060	2.374	2.551	2.673	2.766	2.842	2.959	3.048	3.121	3.208	3.320
30		2.042	2.350	2.522	2.641	2.732	2.805	2.918	3.005	3.075	3.160	3.267
40		2.021	2.321	2.488	2.603	2.690	2.760	2.869	2.952	3.019	3.100	3.203
60		2.000	2.292	2.454	2.564	2.649	2.716	2.821	2.900	2.964	3.041	3.139
						γ = 0.99						
3		5.841	7.127	7.914	8.479	8.919	9.277	9.838	10.269	10.616	11.034	11.559
4		4.604	5.462	5.985	6.362	6.656	6.897	7.274	7.565	7.801	8.087	8.451
5		4.032	4.700	5.106	5.398	5.625	5.812	6.106	6.333	6.519	6.744	7.050
6		3.707	4.271	4.611	4.855	5.046	5.202	5.449	5.640	5.796	5.985	6.250
7		3.500	3.998	4.296	4.510	4.677	4.814	5.031	5.198	5.335	5.502	5.716
8		3.355	3.809	4.080	4.273	4.424	4.547	4.742	4.894	5.017	5.168	5.361
9		3.250	3.672	3.922	4.100	4.239	4.353	4.532	4.672	4.785	4.924	5.103
10		3.169	3.567	3.801	3.969	4.098	4.205	4.373	4.503	4.609	4.739	4.905
11		3.106	3.485	3.707	3.865	3.988	4.087	4.247	4.370	4.470	4.593	4.750
12		3.055	3.418	3.631	3.782	3.899	3.995	4.146	4.263	4.359	4.475	4.625
15		2.947	3.279	3.472	3.608	3.714	3.800	3.935	4.040	4.125	4.229	4.363
20		2.845	3.149	3.323	3.446	3.541	3.617	3.738	3.831	3.907	3.999	4.117
25		2.788	3.075	3.239	3.354	3.442	3.514	3.626	3.713	3.783	3.869	3.978
30		2.750	3.027	3.185	3.295	3.379	3.448	3.555	3.637	3.704	3.785	3.889
40		2.705	2.969	3.119	3.223	3.303	3.367	3.468	3.545	3.607	3.683	3.780
60		2.660	2.913	3.055	3.154	3.229	3.290	3.384	3.456	3.515	3.586	3.676

Source: Reproduced from Hahn and Hendrickson (1971), Biometrika 58, p. 323, with the permission of the author and publisher.

TABLE F1.—Critical values of $t(\alpha;q,\nu)$ for one-sided Dunnett's tests for comparing control against each of q other treatments

	$\alpha = .05$									$\alpha = .01$								
ν \ q	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
5	2.02	2.44	2.68	2.85	2.98	3.08	3.16	3.24	3.30	3.37	3.90	4.21	4.43	4.60	4.73	4.85	4.94	5.03
6	1.94	2.34	2.56	2.71	2.83	2.92	3.00	3.07	3.12	3.14	3.61	3.88	4.07	4.21	4.33	4.43	4.51	4.59
7	1.89	2.27	2.48	2.62	2.73	2.82	2.89	2.95	3.01	3.00	3.42	3.66	3.83	3.96	4.07	4.15	4.23	4.30
8	1.86	2.22	2.42	2.55	2.66	2.74	2.81	2.87	2.92	2.90	3.29	3.51	3.67	3.79	3.88	3.96	4.03	4.09
9	1.83	2.18	2.37	2.50	2.60	2.68	2.75	2.81	2.86	2.82	3.19	3.40	3.55	3.66	3.75	3.82	3.89	3.94
10	1.81	2.15	2.34	2.47	2.56	2.64	2.70	2.76	2.81	2.76	3.11	3.31	3.45	3.56	3.64	3.71	3.78	3.83
11	1.80	2.13	2.31	2.44	2.53	2.60	2.67	2.72	2.77	2.72	3.06	3.25	3.38	3.48	3.56	3.63	3.69	3.74
12	1.78	2.11	2.29	2.41	2.50	2.58	2.64	2.69	2.74	2.68	3.01	3.19	3.32	3.42	3.50	3.56	3.62	3.67
13	1.77	2.09	2.27	2.39	2.48	2.55	2.61	2.65	2.71	2.65	2.97	3.15	3.27	3.37	3.44	3.51	3.56	3.61
14	1.76	2.08	2.25	2.37	2.46	2.53	2.59	2.64	2.69	2.62	2.94	3.11	3.23	3.32	3.40	3.46	3.51	3.56
15	1.75	2.07	2.24	2.36	2.44	2.51	2.57	2.62	2.67	2.60	2.91	3.08	3.20	3.29	3.36	3.42	3.47	3.52
16	1.75	2.06	2.23	2.34	2.43	2.50	2.56	2.61	2.65	2.58	2.88	3.05	3.17	3.26	3.33	3.39	3.44	3.48
17	1.74	2.05	2.22	2.33	2.42	2.49	2.54	2.59	2.64	2.57	2.86	3.03	3.14	3.23	3.30	3.36	3.41	3.45
18	1.73	2.04	2.21	2.32	2.41	2.48	2.53	2.58	2.62	2.55	2.84	3.01	3.12	3.21	3.27	3.33	3.38	3.42
19	1.73	2.03	2.20	2.31	2.40	2.47	2.52	2.57	2.61	2.54	2.83	2.99	3.10	3.18	3.25	3.31	3.36	3.40
20	1.72	2.03	2.19	2.30	2.39	2.46	2.51	2.56	2.60	2.53	2.81	2.97	3.08	3.17	3.23	3.29	3.34	3.38
24	1.71	2.01	2.17	2.28	2.36	2.43	2.48	2.53	2.57	2.49	2.77	2.92	3.03	3.11	3.17	3.22	3.27	3.31
30	1.70	1.99	2.15	2.25	2.33	2.40	2.45	2.50	2.54	2.46	2.72	2.87	2.97	3.05	3.11	3.16	3.21	3.24
40	1.68	1.97	2.13	2.23	2.31	2.37	2.42	2.47	2.51	2.42	2.68	2.82	2.92	2.99	3.05	3.10	3.14	3.18
60	1.67	1.95	2.10	2.21	2.28	2.35	2.39	2.44	2.48	2.39	2.64	2.78	2.87	2.94	3.00	3.04	3.08	3.12
120	1.66	1.93	2.08	2.18	2.26	2.32	2.37	2.41	2.45	2.36	2.60	2.73	2.82	2.89	2.94	2.99	3.03	3.06
∞	1.64	1.92	2.06	2.16	2.23	2.29	2.34	2.38	2.42	2.33	2.56	2.68	2.77	2.84	2.89	2.93	2.97	3.00

Source: Reproduced from C.W. Dunnett, A multiple comparision procedure for comparing several treatments with a control, Journal of the American Statistical Association, vol. 50 (1955).

TABLE F2.—Critical values of $t(\alpha;a,\nu)$ for two-sided Dunnett's tests for comparing control against each of q other treatments

$\alpha = .05$

ν \ q	1	2	3	4	5	6	7	8	9	10	11	12	15	20
5	2.57	3.03	3.29	3.48	3.62	3.73	3.82	3.90	3.97	4.03	4.09	4.14	4.26	4.42
6	2.45	2.86	3.10	3.26	3.39	3.49	3.57	3.64	3.71	3.76	3.81	3.86	3.97	4.11
7	2.36	2.75	2.97	3.12	3.24	3.33	3.41	3.47	3.53	3.58	3.63	3.67	3.78	3.91
8	2.31	2.67	2.88	3.02	3.13	3.22	3.29	3.35	3.41	3.46	3.50	3.54	3.64	3.76
9	2.26	2.61	2.81	2.95	3.05	3.14	3.20	3.26	3.32	3.36	3.40	3.44	3.53	3.65
10	2.23	2.57	2.76	2.89	2.99	3.07	3.14	3.19	3.24	3.29	3.33	3.36	3.45	3.57
11	2.20	2.53	2.72	2.84	2.94	3.02	3.08	3.14	3.19	3.23	3.27	3.30	3.39	3.50
12	2.18	2.50	2.68	2.81	2.90	2.98	3.04	3.09	3.14	3.18	3.22	3.25	3.34	3.45
13	2.16	2.48	2.65	2.78	2.87	2.94	3.00	3.06	3.10	3.14	3.18	3.21	3.29	3.40
14	2.14	2.46	2.63	2.75	2.84	2.91	2.97	3.02	3.07	3.11	3.14	3.18	3.26	3.36
15	2.13	2.44	2.61	2.73	2.82	2.89	2.95	3.00	3.04	3.08	3.12	3.15	3.23	3.33
16	2.12	2.42	2.59	2.71	2.80	2.87	2.92	2.97	3.02	3.06	3.09	3.12	3.20	3.30
17	2.11	2.41	2.58	2.69	2.78	2.85	2.90	2.95	3.00	3.03	3.07	3.10	3.18	3.27
18	2.10	2.40	2.56	2.68	2.76	2.83	2.89	2.94	2.98	3.01	3.05	3.08	3.16	3.25
19	2.09	2.39	2.55	2.66	2.75	2.81	2.87	2.92	2.96	3.00	3.03	3.06	3.14	3.23
20	2.09	2.38	2.54	2.65	2.73	2.80	2.86	2.90	2.95	2.98	3.02	3.05	3.12	3.22
24	2.06	2.35	2.51	2.61	2.70	2.76	2.81	2.86	2.90	2.94	2.97	3.00	3.07	3.16
30	2.04	2.32	2.47	2.58	2.66	2.72	2.77	2.82	2.86	2.89	2.92	2.95	3.02	3.11
40	2.02	2.29	2.44	2.54	2.62	2.68	2.73	2.77	2.81	2.85	2.87	2.90	2.97	3.06
60	2.00	2.27	2.41	2.51	2.58	2.64	2.69	2.73	2.77	2.80	2.83	2.86	2.92	3.00
120	1.98	2.24	2.38	2.47	2.55	2.60	2.65	2.69	2.73	2.76	2.79	2.81	2.87	2.95
∞	1.96	2.21	2.35	2.44	2.51	2.57	2.61	2.65	2.69	2.72	2.74	2.77	2.83	2.91

$\alpha = .01$

ν \ q	1	2	3	4	5	6	7	8	9	10	11	12	15	20
5	4.03	4.63	4.98	5.22	5.41	5.56	5.69	5.80	5.89	5.98	6.05	6.12	6.30	6.52
6	3.71	4.21	4.51	4.71	4.87	5.00	5.10	5.20	5.28	5.35	5.41	5.47	5.62	5.81
7	3.50	3.95	4.21	4.39	4.53	4.64	4.74	4.82	4.89	4.95	5.01	5.06	5.19	5.36
8	3.36	3.77	4.00	4.17	4.29	4.40	4.48	4.56	4.62	4.68	4.73	4.78	4.90	5.05
9	3.25	3.63	3.85	4.01	4.12	4.22	4.30	4.37	4.43	4.48	4.53	4.57	4.68	4.82
10	3.17	3.53	3.74	3.88	3.99	4.08	4.16	4.22	4.28	4.33	4.37	4.42	4.52	4.65
11	3.11	3.45	3.65	3.79	3.89	3.98	4.05	4.11	4.16	4.21	4.25	4.29	4.39	4.52
12	3.05	3.39	3.58	3.71	3.81	3.89	3.96	4.02	4.07	4.12	4.16	4.19	4.29	4.41
13	3.01	3.33	3.52	3.65	3.74	3.82	3.89	3.94	3.99	4.04	4.08	4.11	4.20	4.32
14	2.98	3.29	3.47	3.59	3.69	3.76	3.83	3.88	3.93	3.97	4.01	4.05	4.13	4.24
15	2.95	3.25	3.43	3.55	3.64	3.71	3.78	3.83	3.88	3.92	3.95	3.99	4.07	4.18
16	2.92	3.22	3.39	3.51	3.60	3.67	3.73	3.78	3.83	3.87	3.91	3.94	4.02	4.13
17	2.90	3.19	3.36	3.47	3.56	3.63	3.69	3.74	3.79	3.83	3.86	3.90	3.98	4.08
18	2.88	3.17	3.33	3.44	3.53	3.60	3.66	3.71	3.75	3.79	3.83	3.86	3.94	4.04
19	2.86	3.15	3.31	3.42	3.50	3.57	3.63	3.68	3.72	3.76	3.79	3.83	3.90	4.00
20	2.85	3.13	3.29	3.40	3.48	3.55	3.60	3.65	3.69	3.73	3.77	3.80	3.87	3.97
24	2.80	3.07	3.22	3.32	3.40	3.47	3.52	3.57	3.61	3.64	3.68	3.70	3.78	3.87
30	2.75	3.01	3.15	3.25	3.33	3.39	3.44	3.49	3.52	3.56	3.59	3.62	3.69	3.78
40	2.70	2.95	3.09	3.19	3.26	3.32	3.37	3.41	3.44	3.48	3.51	3.53	3.60	3.68
60	2.66	2.90	3.03	3.12	3.19	3.25	3.29	3.33	3.37	3.40	3.42	3.45	3.51	3.59
120	2.62	2.85	2.97	3.06	3.12	3.18	3.22	3.26	3.29	3.32	3.35	3.37	3.43	3.51
∞	2.58	2.79	2.92	3.00	3.06	3.11	3.15	3.19	3.22	3.25	3.27	3.29	3.35	3.42

Source: Reproduced from C.W. Dunnett, New tables for multiple comparisions with a control, Biometrics 20 (1964), with the permission of the author and the editor.

TABLE G.—Critical values of $\bar{t}(\alpha;p,\nu)$ for testing zero against nonzero dose levels (p = number of nonzero levels)

$\alpha = .05$

ν \ p	1	2	3	4	5	6	7	8	9	10
5	2.02	2.14	2.19	2.21	2.22	2.23	2.24	2.24	2.25	2.25
6	1.94	2.06	2.10	2.12	2.13	2.14	2.14	2.15	2.15	2.15
7	1.89	2.00	2.04	2.06	2.07	2.08	2.08	2.09	2.09	2.09
8	1.86	1.96	2.00	2.01	2.02	2.03	2.04	2.04	2.04	2.04
9	1.83	1.93	1.96	1.98	1.99	2.00	2.00	2.01	2.01	2.01
10	1.81	1.91	1.94	1.96	1.97	1.97	1.98	1.98	1.98	1.98
11	1.80	1.89	1.92	1.94	1.94	1.95	1.95	1.96	1.96	1.96
12	1.78	1.87	1.90	1.92	1.93	1.93	1.94	1.94	1.94	1.94
13	1.77	1.86	1.89	1.90	1.91	1.92	1.92	1.93	1.93	1.93
14	1.76	1.85	1.88	1.89	1.90	1.91	1.91	1.91	1.92	1.92
15	1.75	1.84	1.87	1.88	1.89	1.90	1.90	1.90	1.90	1.91
16	1.75	1.83	1.86	1.87	1.88	1.89	1.89	1.89	1.90	1.90
17	1.74	1.82	1.85	1.87	1.87	1.88	1.88	1.89	1.89	1.89
18	1.73	1.82	1.85	1.86	1.87	1.87	1.88	1.88	1.88	1.88
19	1.73	1.81	1.84	1.85	1.86	1.87	1.87	1.87	1.87	1.88
20	1.72	1.81	1.83	1.85	1.86	1.86	1.86	1.87	1.87	1.87
22	1.72	1.80	1.83	1.84	1.85	1.85	1.85	1.86	1.86	1.86
24	1.71	1.79	1.82	1.83	1.84	1.84	1.85	1.85	1.85	1.85
26	1.71	1.79	1.81	1.82	1.83	1.84	1.84	1.84	1.84	1.85
28	1.70	1.78	1.81	1.82	1.83	1.83	1.83	1.84	1.84	1.84
30	1.70	1.78	1.80	1.81	1.82	1.83	1.83	1.83	1.83	1.83
35	1.69	1.77	1.79	1.80	1.81	1.82	1.82	1.82	1.82	1.83
40	1.68	1.76	1.79	1.80	1.80	1.81	1.81	1.81	1.82	1.82
60	1.67	1.75	1.77	1.78	1.79	1.79	1.80	1.80	1.80	1.80
120	1.66	1.73	1.75	1.77	1.77	1.78	1.78	1.78	1.78	1.78
∞	1.645	1.716	1.739	1.750	1.756	1.760	1.763	1.765	1.767	1.768

TABLE G.—Critical values of $t(\alpha p,v)$ for testing zero against nonzero dose levels (p = number of nonzero levels)—Continued

$\alpha = .01$

ν \ p	1	2	3	4	5	6	7	8	9	10
5	3.36	3.50	3.55	3.57	3.59	3.60	3.60	3.61	3.61	3.61
6	3.14	3.26	3.29	3.31	3.32	3.33	3.34	3.34	3.34	3.35
7	3.00	3.10	3.13	3.15	3.16	3.16	3.17	3.17	3.17	3.17
8	2.90	2.99	3.01	3.03	3.04	3.04	3.05	3.05	3.05	3.05
9	2.82	2.90	2.93	2.94	2.95	2.95	2.96	2.96	2.96	2.96
10	2.76	2.84	2.86	2.88	2.88	2.89	2.89	2.89	2.90	2.90
11	2.72	2.79	2.81	2.82	2.83	2.83	2.84	2.84	2.84	2.84
12	2.68	2.75	2.77	2.78	2.79	2.79	2.79	2.80	2.80	2.80
13	2.65	2.72	2.74	2.75	2.75	2.76	2.76	2.76	2.76	2.76
14	2.62	2.69	2.71	2.72	2.72	2.73	2.73	2.73	2.73	2.73
15	2.60	2.66	2.68	2.69	2.70	2.70	2.70	2.71	2.71	2.71
16	2.58	2.64	2.66	2.67	2.68	2.68	2.68	2.68	2.68	2.69
17	2.57	2.63	2.64	2.65	2.66	2.66	2.66	2.66	2.67	2.67
18	2.55	2.61	2.63	2.64	2.64	2.64	2.65	2.65	2.65	2.65
19	2.54	2.60	2.61	2.62	2.63	2.63	2.63	2.63	2.63	2.63
20	2.53	2.58	2.60	2.61	2.61	2.62	2.62	2.62	2.62	2.62
22	2.51	2.56	2.58	2.59	2.59	2.59	2.60	2.60	2.60	2.60
24	2.49	2.55	2.56	2.57	2.57	2.57	2.58	2.58	2.58	2.58
26	2.48	2.53	2.55	2.55	2.56	2.56	2.56	2.56	2.56	2.56
28	2.47	2.52	2.53	2.54	2.54	2.55	2.55	2.55	2.55	2.55
30	2.46	2.51	2.52	2.53	2.53	2.54	2.54	2.54	2.54	2.54
35	2.44	2.49	2.50	2.51	2.51	2.51	2.51	2.52	2.52	2.52
40	2.42	2.47	2.48	2.49	2.49	2.50	2.50	2.50	2.50	2.50
60	2.39	2.43	2.45	2.45	2.46	2.46	2.46	2.46	2.46	2.46
120	2.36	2.40	2.41	2.42	2.42	2.42	2.42	2.42	2.42	2.43
∞	2.326	2.366	2.377	2.382	2.385	2.386	2.387	2.388	2.389	2.389

Source: Reproduced from D.A.Williams, A test for differences between treatment means when several dose levels are comparedwith a zero dose control, Biometrics 27 (1971), with the permission of the author and the editor.

LIST OF REFERENCES

Alam, K., and Saxena, K. M. L. 1974. On Interval Estimation of a Ranked Parameter. Jour. Roy. Statis. Soc. B 36: 277-283.

Anderson. V. L., and McLean, R. A. 1974. Design of Experiments: A Realistic Approach. Marcel Dekker, Inc., New York.

Arvesen, J. N., and McCabe, G. P., Jr. 1975. Subset Selection Problems for Variances With Applications to Regression Analysis. Jour. Amer. Statis. Assoc. 70: 166-170.

Balaam, L. N. 1963. Multiple Comparisons — A Sampling Experiment. Austral. Jour. Statis. 5: 62-85.

Bancroft, T. A. 1968. Topics in Intermediate Statistical Methods. V. 1. Iowa State Univ. Press, Ames.

Barlow, R. E., and Gupta, S. S. 1969. Selection Procedures for Restricted Families of Probability Distributions. Ann. Math. Statis. 40: 905-934.

Bartholomew, D. J. 1961. Ordered Tests in the Analysis of Variance. Biometrika 48: 325-332.

Bechhofer, R. E. 1968. Single-stage Procedures for Ranking Multiply-Classified Variances of Normal Populations. Technometrics 10: 693-714.

_____ 1969. Optimal Allocation of Observations When Comparing Several Treatments With a Control. *In* Multivariate Analysis-II, P. R. Krishnaiah, ed., pp. 463-473. Academic Press, New York.

_____ Kiefer, J., and Sobel, M. 1968. Sequential Identification and Ranking Procedures. Univ. Chicago Press, Chicago.

_____ Elmaghraby, S., and Morse, N. 1959. A Single-Sample Multiple-Decision Procedure for Selecting the Multinomial Event Which Has the Highest Probability. Ann. Math. Statis. 30: 102-119.

Bernhardson, C. A. 1975. Type I Error Rates When Multiple Comparison Procedures Follow a Significant *F* Test of ANOVA. Biometrics 31: 229-232.

Beyer, W. H., ed. 1968. Handbook of Tables for Probability and Statistics. 2d ed. The Chemical Rubber Co., Cleveland.

Bhargava, R. P., and Srivastava, M. S. 1973. On Tukey's Confidence Intervals for the Contrasts of Means for the Intraclass Correlation Model. Jour. Roy. Statis. Soc. B 35: 147-152.

Bland, R. P., and Bratcher, T. L. 1968. A Bayesian Approach to the Problem of Ranking Binomial Probabilities. SIAM Jour. Appl. Math. 16: 843-850.

Boardman, T. J., and Moffitt, D. R. 1971. Graphical Monte Carlo Type I Error Rates for Multiple Comparison Procedures. Biometrics 27: 738-744.

Bohrer, R. 1967. On Sharpening Scheffe's Bounds. Jour. Roy. Statis. Soc. B 29: 110-114.

Box, G. E. P., and Hunter, J. F. 1958. Experimental Designs for Exploring Response Surfaces. *In* Experimental Designs in Industry. Victor Chew, ed., pp. 138-190. John Wiley and Sons, Inc., New York.

Bradu, D., and Gabriel, K. R. 1974. Simultaneous Statistical Inference on Interactions in Two-Way Analysis of Variance. Jour. Amer. Statis. Assoc. 69: 428-436.

Brown, M. B., and Forsythe, A. B. 1974. The ANOVA and Multiple Comparisons for Data With Heterogeneous Variances. Biometrics 30: 719-724.

Carmer, S. G., and Swanson, M. R. 1971. Detection of Differences Between Means: A Monte Carlo Study of Five Pairwise Multiple Comparisons Procedures. Agron. Jour. 63: 940-945.

Carmer, S. G., and Swanson, M. R. 1973. Evaluation of Ten Pairwise Multiple Comparison Procedures by Monte Carlo Methods. Jour. Amer. Statis Assoc. 68: 66-74.

Chew, V. 1962. Regression Techniques in the Analysis of Variance. Industrial Quality Control. v. 18, No. 12, pp. 1-2.

Chiu, W. K. 1974a. Selecting the m Populations With Largest Means From k Normal Populations With Unknown Variances. Austral. Jour. Statis. 16: 144-147.

_____ 1974b. The Ranking of Means of Normal Populations for a Generalized Selection Goal. Biometrika 61: 579-584.

Cochran, W. G., and Cox, G. M. 1957. Experimental Designs, 2d ed. John Wiley and Co., New York.

Cornell, J. A. 1971. A Review of Multiple Comparison Procedures for Comparing a Set of k Population Means. Soil Crop Sci. Soc. Fla. Proc. 31: 92-97.

Cox, D. R. 1965. A Remark on Multiple Comparison Methods. Technometrics 6: 223-224.

David, H. A. 1956. The Ranking of Variances in Normal Populations. Jour. Amer. Statis. Assoc. 51: 621-626.
_____ 1962. Multiple Decisions and Multiple Comparisons, Chapter 9. *In* Contributions to Order Statistics. Sarhan, A. E., and Greenberg, G. B., ed., John Wiley and Sons, Inc., pp. 144-162, New York.
Davies, O. L., ed. 1956. The Design and Analysis of Industrial Experiments. Oliver and Boyd, Edinburgh.
Dixon, D. O., and Duncan, D. B. 1975. Minimum Bayes Risk *t*-Intervals for Multiple Comparisons. Jour. Amer. Statis. Assoc. 70: 822-831.
Dudewicz, E. J. 1976. Introduction to Statistics and Probability (Ch. 11, Ranking and Selection Procedures). Holt, Rinehart and Winston, New York.
_____ Ramberg, J. S., and Chen, H. J. 1975. New Tables for Multiple Comparisons With a Control (Unknown Variances). Biometrische Zeitschrift 17: 13-26.
Duncan, D. B. 1955. Multiple Range and Multiple F Tests. Biometrics 11: 1-42.
_____ 1957. Multiple Range Tests for Correlated and Heteroscedastic Means. Biometrics 13: 164-176.
_____ 1965. A Bayesian Approach to Multiple Comparisons. Technometrics 7: 171-222.
_____ 1970. Answer to Query #273, Multiple Comparison Methods for Comparing Regression Coefficients. Biometrics 26: 141-143.
_____ 1975. *t* Tests and Intervals for Comparisons Suggested by the Data. Biometrics 31: 339-359.
Dunn, O. J. 1961. Multiple Comparisons Among Means. Jour. Amer. Statis. Assoc. 56: 52-64
_____ 1964. Multiple Comparisons Using Rank Sums. Technometrics 6: 241-252.
_____ and Massey, F. J., Jr. 1965. Estimation of Multiple Contrasts Using *t*-Distributions. Jour. Amer. Statis. Assoc. 60: 573-583.
Dunnett, C. W. 1955. A Multiple Comparisons Procedure for Comparing Several Treatments With a Control. Jour. Amer. Statis. Assoc. 50: 1096-1121.
_____ 1964. New Tables for Multiple Comparisons With a Control. Biometrics 20: 482-491.
_____ 1970. Multiple Comparison Tests (Query #272). Biometrics 26: 139-141.
Eaton, M. L. 1967. Some Optimum Properties of Ranking Procedures. Ann. Math. Statis. 38: 124-137.
Einot, I., and Gabriel, K. R. 1975. A Study of the Powers of Several Methods of Multiple Comparisons. Jour. Amer. Statis. Assoc. 70: 574-583.
Federer, W. T. 1955. Experimental Design, Theory and Application. Macmillan & Co., New York.
_____ 1961. Experimental Error Rates. Amer. Soc. Hort. Sci. Proc. 78: 605-615.
Fienberg, S. E., and Holland, P. W. 1973. Simultaneous Estimation of Multinomial Cell Probabilities. Jour. Amer. Statis. Assoc. 68: 683-691.
Fisher, R. A. 1935. The Design of Experiments. 1st ed. Oliver and Boyd, London.
_____ and Yates, F. 1963. Statistical Tables for Biological, Agricultural, and Medical Research. 6th ed. Oliver and Boyd Ltd., Edinburgh.
Gabriel, K. R. 1964. A Procedure for Testing the Homogeneity of all Sets of Means in Analysis of Variance. Biometrics 20: 459-477.
_____ 1966. Simultaneous Test Procedures for Multiple Comparisons on Categorical Data. Jour. Amer. Statis. Assoc. 61: 1081-1096.
_____ 1968. Simultaneous Test Procedures in Multivariate Analysis of Variance. Biometrika 55: 489-504.
_____ 1969a. Simultaneous Test Procedures — Some Theory of Multiple Comparisons. Ann. Math. Statis. 40: 224-250.
Gabriel, K. R. 1969b. A Comparison of Some Methods of Simultaneous Inference in MANOVA. *In* Multivariate Analysis-II. P. R. Krishnaiah, ed., pp. 67-88. Academic Press, New York.
Games, P. A. 1971. Multiple Comparisons of Means. Amer. Ed. Res. Jour. 8: 531-565.
Gill, J. L. 1973. Current Status of Multiple Comparisons of Means in Designed Experiments. Jour. Dairy Sci. 56: 973-977.
Goodman, L. A. 1965. On Simultaneous Confidence Intervals for Multinomial Proportions. Technometrics 7: 247-254.
Gupta, S. S. 1963. On a Selection and Ranking Procedure for Gamma Populations. Ann. Inst. Statis. Math. 14: 199-216.
_____ 1965. On Some Multiple Decision (Selection and Ranking) Rules. Technometrics 6: 225-245.
_____ and Sobel, M. 1957. On a Statistic Which Arises in Selection and Ranking Problems. Ann. Math. Statis. 28: 957-967.

_____ and Sobel, M. 1958. On Selecting a Subset Which Contains All Populations Better Than a Standard. Ann. Math. Statis 29: 235-244.
_____ and Sobel, M. 1960. Selecting a Subset Containing the Best of Several Binomial Populations. *In* Contribution to Probability and Statistics, ch. 20. Stanford University Press, Stanford.
_____ and Panchapakesan, S. 1971. Contributions to Multiple Decision (Subset Selection) Rules, Multivariate Distribution Theory and Order Statistics. Report No. 71-0218. Aerospace Res. Lab., AFSC, USAF, Wright-Patterson AFB, Ohio.
_____ and Panchapakesan, S. 1972. On a Class of Subset Selection Procedures. Ann. Math. Statis. 43: 814-822.
Hahn, G. J. 1970. Prediction Intervals for a Normal Distribution. Gen. Elec. Co. TIS Rpt. No. 71-C-038. Gen. Elec. Co., Schenectady.
_____ 1972. Simultaneous Prediction Intervals for a Regression Model. Technometrics 14: 203-214.
_____ and Hendrickson, R. W. 1971. A Table of Percentage Points of the Distribution of the Largest Absolute Value of k Student t Variates and its Applications. Biometrika 58: 323-332.
Halperin, M., and Greenhouse, S. W. 1958. A Note on Multiple Comparisons for Adjusted Means in the Analysis of Covariance. Biometrika 45: 256-259.
Harter, H. L. 1957. Error Rates and Sample Sizes for Range Tests in Multiple Comparisons. Biometrics 13: 511-536.
_____ 1960a. Critical Values for Duncan's New Multiple Range Tests. Biometrics 16: 671-685.
_____ 1960b. Tables of Range and Studentized Range. Ann. Math. Statis. 31: 1122-1147.
_____ 1961. Corrected Error Rates for Duncan's New Multiple Range Test. Biometrics 17: 321-324.
_____ 1970. Order Statistics and Their Use in Testing and Estimation. v. l. Tests Based on Range and Studentized Range of Samples from a Normal Population. (Contains updated versions of Harter's Biometrics (1957, 1960, 1961), Technometrics (1961), and AMS (1960) papers.) U.S. Govt. Print. Off., Washington, D.C.
_____ 1970. Multiple comparison procedures for interactions. Amer. Statis. 24: 30-32.
Hartigan, J. A. 1975. Clustering Algorithms. John Wiley and Co., Inc., New York.
Hartley, H. O. 1955. Some Recent Developments in Analysis of Variance. Communications on Pure and Applied Mathematics 8: 47-72.
Hochberg, Y. 1975. An Extension of the T-Method to General Unbalanced Models of Fixed Effects. Jour. Roy. Statis. Soc. B 37: 426-433.
_____ 1976. A Modification of the T-Method of Multiple Comparisons for a One-Way Layout with Unequal Variances. Jour. Amer. Statis. Assoc. 71: 200-203.
_____ and Quade, D. 1975. One-Sided Simultaneous Confidence Bounds on Regression Surfaces With Intercepts. Jour. Amer. Statis. Assoc. 70: 889-891.
Hoel, D., and Sobel, M. 1972. Comparisons of Sequential Procedures for Selecting the Best Binomial Population. *In* Sixth Berkeley Symposium Math. Statis. Probability Proc., v. 4, pp. 53-69.
Hollander, M., and Wolfe, D. A. 1973. Nonparametric Statistical Methods. John Wiley and Co., New York.
Jensen, D. R. 1976. The Comparison of Several Response Functions With a Standard. Biometrics 32: 51-59.
_____ and Jones, M. Q. 1969. Simultaneous Confidence Intervals for Variances. Jour. Amer. Statis. Assoc. 64: 324-332.
John, P. W. M. 1971. Statistical Design and Analysis of Experiments. The MacMillan Company, New York.
Johnson, D. E. 1976. Some New Multiple Comparison Procedures for the Two-Way AOV Model With Interaction. Biometrics 32: 929–934.
Jolliffe, I. T. 1975. Cluster Analysis as a Multiple Comparison Method. *In* Applied Statistics. R. P. Gupta, ed. North-Holland Pub. Co., New York.
Kappenman, R. F. 1972. A Note on Selection of the Greatest Exceedance Probability. Technometrics 14: 219-222.
Keselman, H. J., Toothaker, L. E., and Shooter, M. 1975. An Evaluation of Two Unequal n_k forms of the Tukey Multiple Comparison Statistic. Jour. Amer. Statis. Assoc. 70: 584-587.
Keuls, M. 1952. The Use of the "Studentized Range" in Connection With an Analysis of Variance. Euphytica 1: 112-122.
Kirk, R. E. 1968. Experimental Design Procedures for the Behavioral Sciences. Brooks/Cole, Belmont.

Kramer, C. Y. 1956. Extension of Multiple Range Tests to Group Means With Unequal Numbers of Replications. Biometrics 12: 309-310.

_____ 1957. Extension of Multiple Range Tests to Group Correlated Adjusted Means. Biometrics 13: 13-18.

_____ 1972. A First Course in Methods of Multivariate Analysis. Va. Polytech. Inst. State Univ., Blacksburg.

Krishnaiah, P. R. 1969. Simultaneous Test Procedures Under General MANOVA Models. *In* Multivariate Analysis-II, P. R. Krishnaiah, ed., pp. 121-144. Academic Press, New York.

Kuiper, F. K., and Fisher, L. 1975. A Monte Carlo Comparison of Six Clustering Procedures. Biometrics 31: 777-784.

Kurtz, T. E., Link, R. F., Tukey, J. W., and Wallace, D. L. 1965. Short-Cut Multiple Comparisons for Balanced Single and Double Classifications: Part 1, Results. Technometrics 7: 95-169.

LeClerg, E. L. 1957. Mean Separation by the Functional Analysis of Variance and Multiple Comparisons, U.S. Dept. Agr., Agr. Res. Serv., ARS 20-3. (Reprinted July 1970.)

Leonard, T. 1972. Bayesian Methods for Binomial Data. Biometrika 59: 581-589.

Levy, K. J. 1975a. An Empirical Comparison of Several Multiple Range Tests for Variances. Jour. Amer. Statis. Assoc. 70: 180-183.

_____ 1975b. A Multiple Range Procedure for Correlated Variances in a Two-Way Classification. Biometrics 31: 243-246.

Little, T. M., and Hills, F. J. 1972. Statistical Methods in Agricultural Research. Univ. Calif., Agr. Ext. Serv., Davis.

Marriott, F. H. C. 1971. Practical Problems in a Method of Cluster Analysis. Biometrics 27: 501-514.

McCool, J. I. 1975. Multiple Comparisons for Weibull Parameters. IEEE Transactions on Reliability R-24: 186-192.

McDonald, B. J., and Thompson, W. A., Jr. 1967. Rank Sum Multiple Comparisons in One- and Two-Way Classifications. Biometrika 54: 487-497.

Mead, R., and Pike, D. J. 1975. A Review of Response Surface Methodology From a Biometrics Viewpoint. Biometrics 31: 803-852.

Miller, R. G., Jr. 1966. Simultaneous Statistical Inference. McGraw-Hill Book Co., New York.

Morrison, D. F. 1967. Multivariate Statistical Methods. McGraw-Hill Book Co., New York.

Myers, R. H. 1971. Response Surface Methodology. Allyn and Bacon, Inc., Boston.

Nair, K. R. 1948. The Studentized Form of the Extreme Mean Square Test in the Analysis of Variance. Biometrika 35: 16-31.

Newman, D. 1939. The Distribution of the Range in Samples From a Normal Population, Expressed in Terms of an Independent Estimate of Standard Deviation. Biometrika 31: 20-30.

Ofosu, J. B. 1975. A Two-Stage Minimax Procedure for Selecting the Normal Population With the Small Variance. Jour. Amer. Statis. Assoc. 70: 171–174.

O'Neill, R., and Wetherill, G. B. 1971. The Present State of Multiple Comparison Methods. Jour. Roy. Statis. Soc. 70 :171-174.

Patel, J. K. 1976. Ranking and Selection of IFR Populations Based on Means. Jour. Amer. Statis. Assoc. 71: 143-146.

Paulson, E. 1962. A Sequential Procedure for Comparing Several Experimental Categories With a Standard or Control. Ann. Math. Statis. 33: 438-443.

_____ 1964. A Sequential Procedure for Selecting the Population with the Largest Mean From K Normal Populations. Ann. Math. Statis. 35: 174-180.

_____ 1967. Sequential Procedures for Selecting the Best One of Several Binomial Populations. Ann. Math. Statis. 38: 117-123.

Pearson, E. S., and Hartley, H. O. 1966. Biometrika Tables for Statisticians. V. 1, 3d ed. Cambridge Univ. Press, London.

Peng, K. C. 1967. The Design and Analysis of Scientific Experiments. Addison-Wesley Pub. Co., Inc., Reading.

Petrinovich, L. F., and Hardyck, C. D. 1969. Error Rates for Multiple Comparison Methods. Psychol. Bul. 71: 43-54.

Puri, M. L., and Puri, P. S. 1969. Multiple Decision Procedures Based on Ranks for Certain Problems in Analysis of Variance. Ann. Math. Statis. 40: 619-632.
Ramachandran, K. V. 1956. Contributions to simultaneous confidence interval estimation. Biometrics 12: 51-56.
Reading, J. C. 1975. A Multiple Comparison Procedure for Classifying All Pairs out of K Means as Close or Distant. Jour. Amer. Statis. Assoc. 70: 832-838.
Reiersøl, O. 1961. Linear and Non-Linear Multiple Comparisons in Logit Analysis. Biometrika 48: 359-365. Corrigenda, Biometrika 49: 284.
Rhyne, A. L., and Steel, R. G. D. 1965. Tables for a Treatments Versus Control Multiple Comparisons Sign Test. Technometrics 7: 293-306.
_____ and Steel, R. G. D. 1967. A Multiple Comparisons Sign Test: All Pairs of Treatments. Biometrics 23: 539-549.
Rizvi, M. H., Sobel, M., and Woodworth, G. C. 1968. Nonparametric Ranking Procedures for Comparisons With a Control. Ann. Math. Statis. 39: 2075-2093.
_____ 1971. Some Selection Problems Involving Folded Normal Distributions. Technometrics 13: 355-369.
Robbins, H., Sobel, M., and Starr, N. 1968. A Sequential Procedure for Selecting the Largest of K Means. Ann. Math. Statis. 39: 88-92.
Robson, D. S. 1961. Multiple Comparisons With a Control in Balanced Incomplete Block Designs. Technometrics 3: 103-105.
Ryan, T. A. 1959. Multiple Comparisons in Psychological Research. Psychol. Bul. 56: 26-47.
_____ 1960. Significance Tests for Multiple Comparison of Proportions, Variances, and Other Statistics. Psychol. Bul. 57: 318-328.
Ryan, T. A., Jr., and Antle, C. E. 1976. A Note on Gupta's Selection Procedure. Jour. Amer. Statis. Assoc. 71: 140-142,
Santner, T. J. 1975. A Restricted Subset Selection Approach to Ranking and Selection Problems. Ann. Stat. 3: 334-349.
Saxena, K. M. L. 1976. A Single-Sample Procedure for Estimation of the Largest Mean. Jour. Amer. Statis. Assoc. 71: 147-148.
Schafer, W. D., and MacReady, G. B. 1975. A Modification of the Bonferroni Procedure on Contrasts Which Are Grouped Into Internally Independent Sets. Biometrics 31: 227-228.
Scheffé, H. 1953. A Method for Judging All Contrasts in the Analysis of Variance. Biometrika 40: 87-104.
_____ 1959. The Analysis of Variance. John Wiley and Sons, Inc., New York.
Scott, A. J., and Knott, M. 1974. A Cluster Analysis Method for Grouping Means in the Analysis of Variance. Biometrics 30: 507-512.
Seeger, P. 1966. Variance Analysis of Complete Designs. Almqvist and Wiksell, Stockholm.
Sen, P. K. 1969. A Generalization of the T-Method of Multiple Comparisons. Jour. Amer. Statis. Assoc. 64: 290-295.
_____ 1969. On Nonparametric T-Method of Multiple Comparisons for Randomized Blocks. Ann. Inst. Statis. Math. 21: 329-333.
Sherman, E. 1965. A Note on Multiple Comparisons Using Rank Sums. Technometrics 6: 255-256.
Siotani, M. 1964. Interval Estimates for Linear Combinations of Means. Jour. Amer. Statis. Assoc. 59: 1141-1164.
Slivka, J. 1970. A One Sided Nonparametric Multiple Comparison Control Percentile Tests: Treatments Versus Control. Biometrika 57: 431-438.
Sobel, M. 1969. Selecting a Subset Containing at Least One of the T Best Populations. *In* Multivariate Analysis-II. P. R. Krishnaiah, ed. pp. 515-539. Academic Press, New York.
_____ and Tong, Y. L. 1971. Optimal Allocation of Observations for Partitioning a Set of Normal Populations in Comparison With a Control. Biometrika 58: 177-181.
Spjøtvoll, E. 1972. Multiple Comparisons of Regression Functions. Ann. Math. Statis. 72: 1076-1088.
_____ 1972. Joint Confidence Intervals for All Linear Functions of Means in the One-Way Layout With Unknown Group Variances. Biometrika 59: 683-685.
_____ and Stoline, M. R. 1973. An Extension of the T-Method of Multiple Comparison to Include the Cases With Unequal Sample Sizes. Jour. Amer. Statis. Assoc. 68: 975-978.

Steel, R. G. D. 1959. A Multiple Comparison Rank Sum Test: Treatments Versus Control. Biometrics 15: 560-572.

_____ 1961. Some Rank Sum Multiple Comparisons Tests. Biometrics 17: 539-552.

_____ and Torrie, J. H. 1960. Principles and Procedures of Statistics. McGraw-Hill, New York.

Tarone, R. E. 1976. Simultaneous Confidence Ellipsoids in the General Linear Model. Technometrics 18: 85-87.

Taylor, R. J., and David, H. A. 1962. A Multi-Stage Procedure for the Selection of the Best of Several Binomial Populations. Jour. Amer. Statis. Assoc. 57: 785-796.

Thigpen, C. C., and Paulson, A. S. 1974. A Multiple Range Test for Analysis of Covariance. Biometrika 61: 479-484.

Thomas, D. A. H. 1973. Multiple Comparisons Among Means — A Review. Statistician 22: 16-42.

_____ 1974. Error Rates in Multiple Comparisons Among Means — Results of a Simulation Exercise. Jour. Roy. Statis. Soc. C 23: 284-294.

Tobach, E., Smith, M., Rose, G., and Richter, D. 1967. A Table for Rank Sum Multiple Paired Comparisons. Technometrics 9: 561-567.

Tong, Y. L. 1970. Multi-Stage Interval Estimation of the Largest Mean of K Normal Populations. Jour. Roy. Statis. Soc. B 32: 272-277.

Trawinski, B. J., and David, H. A. 1963. Selection of the Best Treatment in a Paired-Comparison Experiment. Ann. Math. Statis. 34: 75-94.

Tukey, J. W. 1949. Comparing Individual Means in the Analysis of Variance. Biometrics 5: 99-114.

_____ 1951. Quick- and-dirty Methods in Statistics, Part 2. Simple Analyses for Standard Designs. Amer. Soc. Qual. Control, 5th Ann. Conv. Trans. pp. 189-197.

_____ 1953a. Some Selected Quick and Easy Methods of Statistical Analysis. Trans. N.Y. Acad. Sci. (2) 16: 88-97.

_____ 1953b. The Problem of Multiple Comparisons. Unpublished Dittoed Notes, Princeton Univ., 396 pp.

_____ 1960. Conclusions vs. Decisions. Technometrics 2: 423-433.

Ury, H. K. 1976. A Comparison of Four Procedures for Multiple Comparisons Among Means (Pairwise Contrasts) for Arbitrary Sample Sizes. Technometrics 18: 89-97.

Verhagen, A. M. W. 1963. The "Caution Level" in Multiple Tests of Significance. Austral. Jour. Statis. 5: 41-48.

Wackerly, D. D. 1975. An Alternative Approach to the Problem of Selecting the Best of K Populations. Technical Report #91. Univ. Fla. Dept. Statis., Gainesville.

Waldo, D. R. 1976. An Evaluation of Multiple Comparison Procedures. Jour. Animal Sci. 42: 539-544.

Waller, R. A., and Duncan, D. B. 1969 and 1972. A Bayes Rule for the Symmetric Multiple Comparison Problem. Jour. Amer. Statis. Assoc. 64: 1484-1503, and Corrigenda 67: 253-255.

Weatherill, G. B., and Ofosu, J. B. 1974. Selection of the Best of K Normal Populations. Jour. Roy. Statis. Soc. C 23: 253-277.

Williams, D. A. 1971. A Test for Differences Between Treatment Means When Several Dose Levels Are Compared With a Zero Dose Control. Biometrics 27: 103-117.

_____ 1972. The Comparison of Several Dose Levels With a Zero Dose Control. Biometrics 28: 519-531.

Wynn, H. P., and Bloomfield, P. 1971. Simultaneous Confidence Bands in Regression Analysis. Jour. Roy. Statis. Soc. B 33: 202-217.

* U.S. GOVERMENT PRINTING OFFICE: 1978 0-280-931/SEA-5